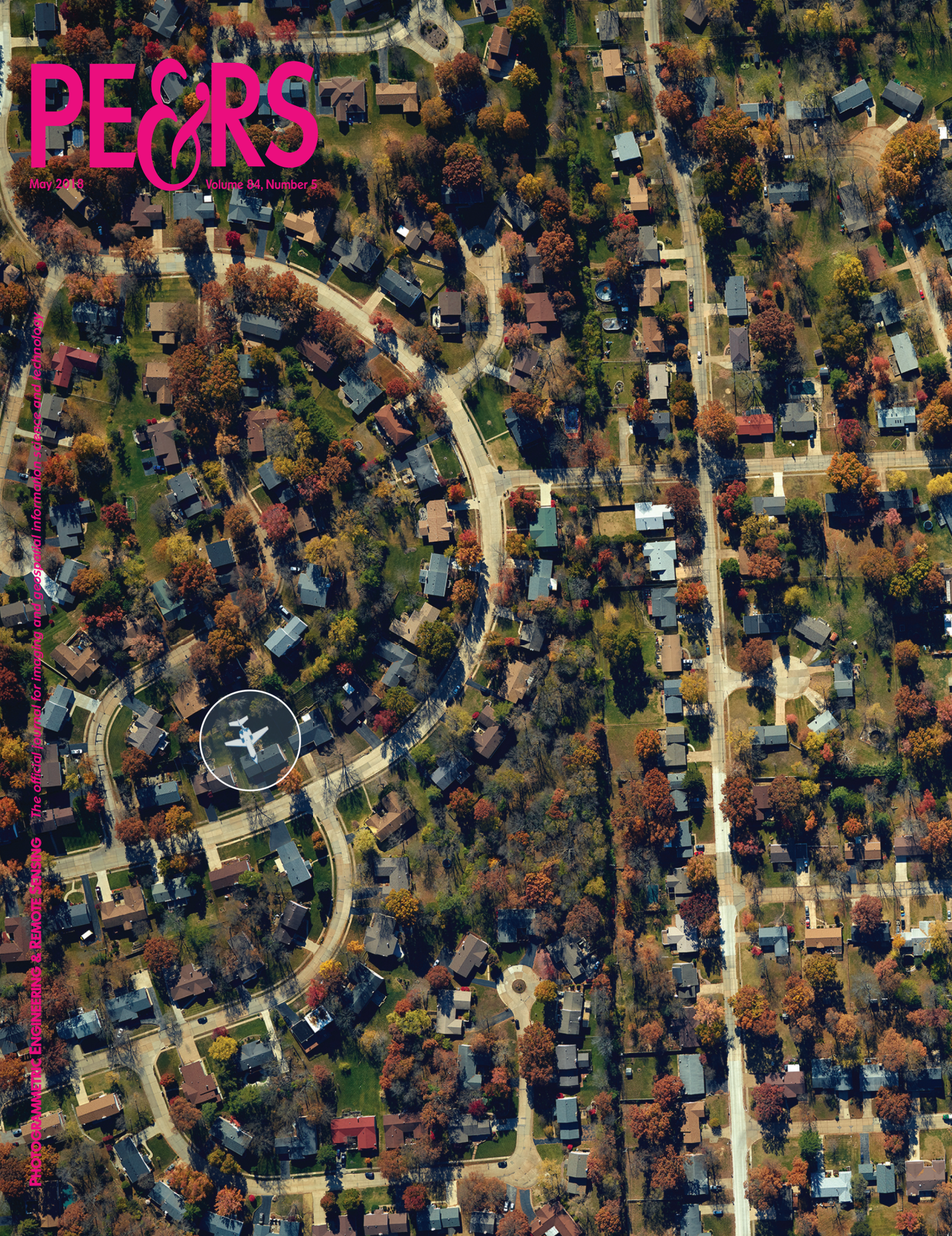
PE&RS

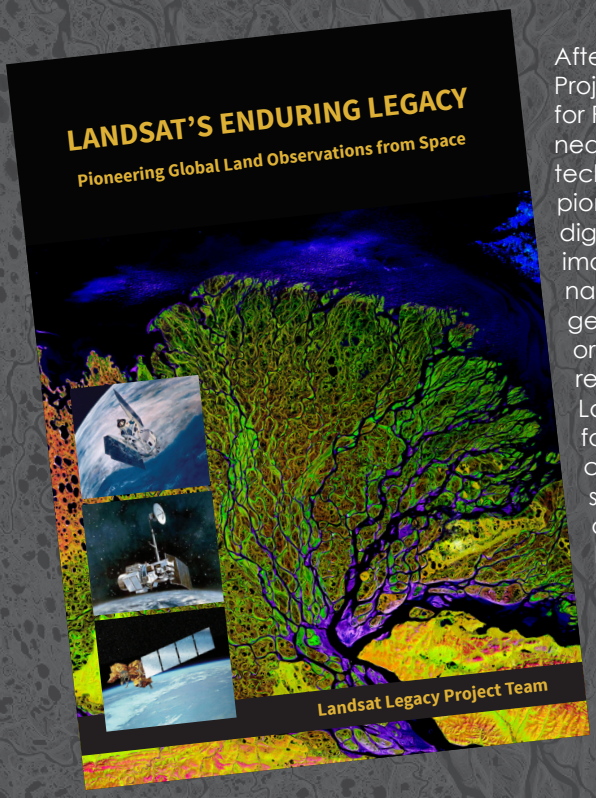
May 2018 Volume 84, Number 5



PHOTOGRAMMETRIC ENGINEERING & REMOTE SENSING *The official journal for imaging and geospatial information science and technology*

Landsat's Enduring Legacy

Pioneering Global Land Observations from Space



After more than 15 years of research and writing, the Landsat Legacy Project Team is about to publish, in collaboration with the American Society for Photogrammetry and Remote Sensing (ASPRS), a seminal work on the nearly half-century of monitoring the Earth's lands with Landsat. Born of technologies that evolved from the Second World War, Landsat not only pioneered global land monitoring but in the process drove innovation in digital imaging technologies and encouraged development of global imagery archives. Access to this imagery led to early breakthroughs in natural resources assessments, particularly for agriculture, forestry, and geology. The technical Landsat remote sensing revolution was not simple or straightforward. Early conflicts between civilian and defense satellite remote sensing users gave way to disagreements over whether the Landsat system should be a public service or a private enterprise. The failed attempts to privatize Landsat nearly led to its demise. Only the combined engagement of civilian and defense organizations ultimately saved this pioneer satellite land monitoring program. With the emergence of 21st century Earth system science research, the full value of the Landsat concept and its continuous 45-year global archive has been recognized and embraced. Discussion of Landsat's future continues but its heritage will not be forgotten.

The pioneering satellite system's vital history is captured in this notable volume on Landsat's Enduring Legacy.

Landsat Legacy Project Team

Samuel N. Goward
Darrel L. Williams
Terry Arvidson
Laura E. P. Rocchio
James R. Irons
Carol A. Russell
Shaida S. Johnston

Landsat's Enduring Legacy

Hardback. 2017, ISBN 1-57083-101-7

Student	$60*
Member	$80*
Non-member	$100*

* Plus shipping

Order online at

www.asprs.org/landsat

ANNOUNCEMENT

The Polis Center at IUPUI announced today the appointment of James I. Sparks as Director of Geoinformatics. A highly-experienced geospatial information professional, Sparks has spent the majority of his career working with geospatial information. He comes to Polis after serving as Indiana's first Geographic Information Officer which entailed coordinating the statewide geospatial efforts and integrating, creating, and distributing geospatial data. Previously, Sparks was integral to the development of the Indianapolis Mapping and Geographic Infrastructure System (IMAGIS), serving as project manager for the data conversion component. This substantial effort converted paper and digital data into GIS layers to create a geographic information system for Marion County, Indiana. Upon completion, IMAGIS was recognized as noteworthy for its size, complexity, and the level of benefit that it delivered.

"This is not my first time to be involved with The Polis Center," Sparks said. "I worked closely with Dr. Bodenhamer and others at the Center in the early 2000s developing geospatial projects characterized by having both a practical and applied orientation -- the same project model still used by The Polis Center today. I am excited about a return to an academic setting while reuniting with great friends and colleagues to help create a spatially-enabled Indiana that is well-positioned to support emerging efforts like 'smart cities' and the Internet of Things."

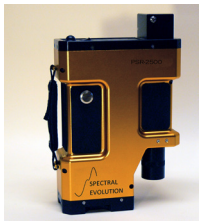
David Bodenhamer, Executive Director of The Polis Center, said "We are delighted to have Jim Sparks as a key member of our team. He has led Indiana's efforts to become a spatially-enabled state, for which he has received national recognition. With his help, we look forward to developing even more opportunities for effective university-government-community partnerships to use spatial information to improve the standard-of-living and enhance the quality of life for Hoosiers."

Jim's professional affiliations include serving as state representative of the National States Geographic Information Council and is a founding member of the Indiana Geographic Information Council (IGIC). Jim earned both a M.S. degree in management and a B.S. degree in business administration from Indiana Wesleyan University.

As a self-funded research unit of the IU School of Liberal Arts at IUPUI, the Polis Center collaborates to create innovative place-based solutions that lead to healthier and more resilient communities. It does that by creating actionable information, developing creative collaborations, doing place-based research, and employing technology effectively to enhance the capacity of communities to respond meaningfully to change. The approach of The Polis Center is practical, applied, and entrepreneurial. It works collaboratively and often serves as the nexus among diverse community-based organizations, government agencies, educational institutions, arts and cultural organizations, businesses, charitable endowments, and faith-based organizations. The Center is committed to linking university and community expertise and to the smart use of advanced technologies to help solve problems and help communities take advantage of opportunities. Geospatial technologies, especially GIS, are its preferred technical tools because of their unique ability to integrate and visualize information by location. The Center uses these tools to develop and analyze data for communities and then involves local experts in helping to understand what the results mean. In doing so, The Polis Center helps communities for more productive decision-making. In all of the sectors in which it works, it has earned a national reputation as a dynamic urban-centered, learning environment with highly professional staff who excel in local experts in helping to understand what the results mean. In doing so, The Polis Center helps communities for more productive decision-making. In all of the sectors in which it works, it has earned a national reputation as a dynamic urban-centered, learning environment with highly professional staff who excel in partnerships, real-world application, and winning solutions for the communities.

PRODUCTS

Spectral Evolution introduces the newly designed PSR-2500 – a full range field spectroradiometer designed to meet research needs and budgets.



The PSR-2500 picks up its new design and construction from Spectral Evolution's leading remote sensing spectroradiometer, the PSR+. This includes an anodized aluminum unibody chassis with integrated heat dispersion channels that's good looking and rugged. With no moving parts, the PSR-2500 is ideal for field research applications. In addition, the instrument has been upgraded to provide improved spectral resolution; 3.5nm @ 700nm, 20nm @ 1500nm, and 18nm @ 2100nm

In addition to improved resolution, the new updated PSR-2500 delivers high sensitivity for better field measurements; ≤0.8x10-9 W/cm2/nm/sr @400nm, ≤1.5x10-9 W/cm2/nm/sr @1500nm, and ≤1.8x10-9 W/cm2/nm/sr @2100nm

The PSR-2500 is field ready, weighing in at 7.3 lbs (3.3 kg), powered by a rechargeable lithium ion battery (two are included with the instrument) and a backpack or shoulder strap for easy field mobility. The PSR-2500 can be used with a range of direct attach lenses or a fiberoptic cable with field-of-view (FOV) lenses: 4°, 8° or 14°, a 25° fiber optic diffuser or integrating sphere. Other accessories include our pistol grip, choice of

a 3mm or 10mm contact probe, desktop probe, and our unique leaf clip that keeps the source of illumination away from your sample. The PSR-2500 can store 1000 scans internally and has an LCD screen for running its DARWin SP Data Acquisition software or use the optional GETAC minicomputer to record digital images, voice notes, GPS and altimeter readings and tag that data to the scans. All files are ASCII format for easy use with 3rd party analysis software.

The PSR-2500 is well-suited for a range of remote sensing applications, including; ground truthing satellite or flyover data, radiance and irradiance measurement, crop and soil studies, forestry and canopy studies, atmospheric research, plant species identification, agricultural analysis, and geological remote sensing.

For more information on the PSR-2500, or any of our remote sensing systems, visit: http://www.spectralevolution.com/portable_spectroradiometer_remote_sensing.html.

Spectral Evolution's SR-4500A is designed for applications that require precise, stable, repeatable performance. Using all thermoelectrically cooled photodiode arrays, the SR-4500A spectroradiometer is built specifically for the performance demands of radiometric calibration transfer. With the SR-4500A, you can take the calibration process where it's needed with the following advanced features:

- Drift stability of 0.1% which delivers greater accuracy for long-term stability of integrating spheres
- Stability is achieved through heating and cooling thermal management features
- A temperature controller maintains the instrument housing at a stable temperature along with the individually temperature stabilitzed detector arrays
- All temperatures are integrated into DARWin software readout for monitoring

The SR-4500A was built to meet the exacting needs of customers for measuring temporal stability. The SR-4500A also features:

- Spectral range of 350-2500nm
- 512 element TE cooled silicon photodiode array (350-1000nm)
- 256 element TE cooled extended In GaAs array (1000-1900nm)
- 256 element TE cooled extended In GaAs array (1900-2500nm)

In addition to stability, the SR-4500A also provides superior Noise Equivalence Radiance capabilities:

- Noise Equivalence Radiance (with 1.2 meter fiber optic)
  - 0.2x10-9 W/cm2/nm/sr @ 400nm
  - 0.2x10-9 W/cm2/nm/sr @ 700nm
  - 0.9x10-9 W/cm2/nm/sr @ 900nm

For more information on the SR-4500A, please contact: Maurice.kashdan@spectralevolution.com.

EVENTS

The conference program for **SPAR 3D Expo & Conference** has been finalized. The program, organized by conference planners and an Advisory Board comprised of 15 industry experts, includes 24 product previews and 48 sessions made up of keynote presentations, plenaries, 101-tracks, and panels. More than 50 speakers including 7 keynoters will provide industry professionals with an overview of 3D technologies, insight into the newest 3D sensing, 3D processing, and 3D visualization technologies on the market, as well as a glimpse at what's to come.

Combined with an international exhibition of the newest 3D technology from the world's top vendors and numerous networking events, the expo and conference is critical for anyone who needs to stay on top of the latest technology and developments in the rapidly changing 3D technology market. The 2018 edition of SPAR 3D Expo & Conference will take place from June 5-7 at the Anaheim Convention Center, Anaheim, California. SPAR 3D Expo & Conference 2018 will be co-located with the inaugural edition of AEC Next Technology Expo & Conference.

The three-day conference and exhibition kicks off on June 5 with the introduction of the latest products by two dozen manufacturers in brief Product Preview presentations.

Keynote speaker Alexander Menzies, representative of the NASA Jet Propulsion Laboratory, will present The Telescope of Tomorrow, followed by keynote Nathan Pettyjohn, Founder of The VR/AR Association, for his lecture Vision for VR/AR Technologies in the Large Enterprise.

After a brief networking break, the conference program splits into separate tracks for Transportation, VDC, and Autonomous Technologies. Experts in each field will share best practices, case studies, and challenges with the participants.

The second day begins with keynote speakers Ken Sanders of Gensler and Jack Dahlgren of NVIDIA co-presenting Creativity & Constructability in AEC. Numerous breakout sessions organized by thematic track will follow the keynote including; AEC & VDC, Mining & Aggregates, 101: Selecting 3D Tech for Reality Capture, Industrial Facilities /Asset Management, Security, Law Enforcement, Forensics, 101: Should your 3D be In-Housed or Outsourced?, Surveying & Mapping, Process, Power, Utilities, and 101: Contracting – RFQ Tips & Pitfalls to Avoid.

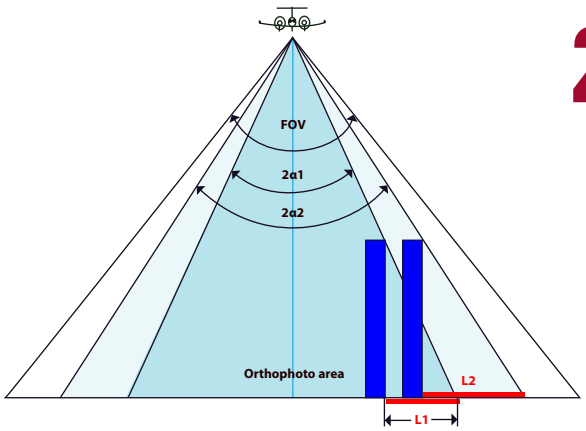
The final day of the conference starts with the two more keynote presentations. First David Wilson, Principal Vice President - Chief Innovation Officer of Bechtel Corporation presents Intelligent Design for Collaborative Constructability. And Atul Khanzode, Chief Information Officer - DPR Construction will follow with his keynote presentation Integrated Projects – A Vision for Future.

Following day three's keynote presentations will be the final

continued on pg 233

PE&RS
PHOTOGRAMMETRIC ENGINEERING & REMOTE SENSING
The official journal for imaging and geospatial information science and technology
May 2018 Volume 84 Number 5

FEATURES



COLUMNS

ANNOUNCEMENTS

DEPARTMENTS

PEER-REVIEWED ARTICLES

See the Cover Description on Page 232

COVER DESCRIPTION

The image showcased on the cover was captured with the Phase One Industrial 190MP Aerial system. iXU-RS1900 with built-in dual RS 90mm lenses, a solution delivering large format functionality.

Captured in the USA, the image was taken with a flight at an altitude of 4,490ft (1,369m) and at 125kts speed, a GSD of 7cm was achieved. The camera's settings were as follow: ISO 100, shutter speed 1/1250s, aperture f/5.6.

Flying at speed of 125 kts, the passing jet aircraft below, who was flying in the opposite direction, can be seen clearly, negating the need for conventional FMC.

The innovative 190MP aerial system, in cooperation with leading partners stabilized mount design, positioning and Flight Management Systems, demonstrated excellent data results that offer faster, more detailed, large format coverage.

Luis Viveros, Americas Technical Manager for Phase One, expressed "The power and quality that comes from such a compact system will save our customers time and improve data collection efficiency, along with the compatibility it offers for a wide range of flying platforms. The resulting data showed images with greater area coverage and high spatial resolution. Ground footprints of 230 acres per frame were collected achieving 7cm GSD, at a flight height of 4,500ft, using a single engine Cessna 206" said Viveros. "Also, the system proved the ability to acquire images with a GSD of 2cm at a flight height of 1,300ft, and still be able to trigger fast enough to obtain 80% of overlap, with a capture rate of less than one second." This capacity is the result of the combination of the high stereoscopic accuracy introduced with the large inline FOV of the new system, and the Reliance Shutter technology already present in Phase One Industrial aerial cameras.

At the heart of the fully integrated 190MP Aerial System is the iXU-RS1900 camera. It features two CMOS sensors and two 90mm lenses for capturing RGB information. Key imaging attributes include a small pixel size (4.6 µm), large image area (16,470 x 11,540), high image capture rate of 0.6 sec and exposure time of up to 1/2000 sec. The Gyro Stabilized Camera mount used was the SOMAG DSM400, with a low weight of 14 kg and high payload of 35 kg, the mount provides optimal stabilization of the system and allows a high-efficient and precise image capturing at any flight conditions.

An optional 4-band configuration, adding a 50 mm lens for capturing NIR information, provides 4-band (RGB, NIR) imagery. Integrated iX Capture software automatically generates distortion-free images and automatically performs an accurate matching of the NIR and RGB images. The Phase One 190MP 4-Band is offered with two GNSS-Inertial Position and Orientation System options, Applanix POS AVX 210 or POS AV 510.

Designed with input from engineers and leading experts in the photogrammetric field to address a wide variety of challenging aerial applications, the enhanced productivity and expanded coverage makes it an ultimate solution for diverse aerial survey applications such as mapping, 3D City modeling, remote sensing, precision agriculture, disaster management and monitoring.

The 190MP Aerial System presents an attractive alternative to other traditionally expensive large format cameras.

PHOTOGRAMMETRIC ENGINEERING & REMOTE SENSING

THE IMAGING & GEOSPATIAL INFORMATION SOCIETY

JOURNAL STAFF

Publisher ASPRS
Editor-In-Chief Alper Yilmaz
Technical Editor Michael S. Renslow
Assistant Editor Jie Shan
Assistant Director — Publications Rae Kelley
Electronic Publications Manager/Graphic Artist Matthew Austin

Photogrammetric Engineering & Remote Sensing is the official journal of the American Society for Photogrammetry and Remote Sensing. It is devoted to the exchange of ideas and information about the applications of photogrammetry, remote sensing, and geographic information systems. The technical activities of the Society are conducted through the following Technical Divisions: Geographic Information Systems, Photogrammetric Applications, Lidar, Primary Data Acquisition, Professional Practice, and Remote Sensing Applications. Additional information on the functioning of the Technical Divisions and the Society can be found in the Yearbook issue of *PE&RS*.

Correspondence relating to all business and editorial matters pertaining to this and other Society publications should be directed to the American Society for Photogrammetry and Remote Sensing, 425 Barlow Place, Suite 210, Bethesda, Maryland 20814-2144, including inquiries, memberships, subscriptions, changes in address, manuscripts for publication, advertising, back issues, and publications. The telephone number of the Society Headquarters is 301-493-0290; the fax number is 301-493-0208; web address is www.asprs.org.

PE&RS. *PE&RS* (ISSN0099-1112) is published monthly by the American Society for Photogrammetry and Remote Sensing, 425 Barlow Place, Suite 210, Bethesda, Maryland 20814-2144. Periodicals postage paid at Bethesda, Maryland and at additional mailing offices.

SUBSCRIPTION. For the 2018 subscription year, ASPRS is offering two options to our *PE&RS* subscribers — an e-Subscription and the print edition. E-subscribers can plus-up their subscriptions with printed copies for a small additional charge. Print subscriptions are on a calendar-year basis that runs from January through December. Electronic subscriptions run for twelve months on an anniversary basis. We recommend that customers who choose both e-Subscription and print (e-Subscription + Print) renew on a calendar-year basis. The new electronic subscription includes access to ten years of digital back issues of *PE&RS* for online subscribers through the same portal at no additional charge. Please see the Frequently Asked Questions about our journal subscriptions.

The rate of the e-Subscription (digital) Site License Only for USA and Foreign: $1000.00; e-Subscription (digital) Site License Only for Canada*: $1049.00; Special Offers: e-Subscription (digital) Plus Print for the USA: $1,365.00; e-Subscription (digital) Plus Print Canada*: $1,424.00; e-Subscription (digital) Plus Print Outside of the USA: $1,395.00; Printed-Subscription Only for USA: $1065.00; Printed-Subscription Only for Canada*: $1,124.00; Printed-Subscription Only for Other Foreign: $1,195.00. *Note: e-Subscription/Printed-Subscription Only/e-Subscription Plus Print for Canada include 5% of the total amount for Canada's Goods and Services Tax (GST #135123065).

POSTMASTER. Send address changes to *PE&RS*, ASPRS Headquarters, 425 Barlow Place, Suite 210, Bethesda, Maryland 20814-2144. CDN CPM #(40020812)

MEMBERSHIP. Membership is open to any person actively engaged in the practice of photogrammetry, photointerpretation, remote sensing and geographic information systems; or who by means of education or profession is interested in the application or development of these arts and sciences. Membership is for one year, with renewal based on the anniversary date of the month joined. Membership Dues include a 12-month electronic subscription to *PE&RS*. Or you can receive the print copy of *PE&RS* Journal which is available to all member types for an additional fee of $60.00 USA and or $75.00 for international shipping. Subscription is part of membership benefits and cannot be deducted from annual dues. Dues for ASPRS Members outside of the U.S. will now be the same as for members residing in the U.S. Annual dues for Regular members (Active Member) is $150; for Student members it is $50 for USA and Canada; $60 for Other Foreign. A tax of 5% for Canada's Goods and Service Tax (GST #135123065) is applied to all members residing in Canada.

continued from pg 230

breakout sessions organized by thematic track including; Civil Infrastructure, Unique Applications (Insurance & Real Estate; Entertainment & Gaming), Attendees – Bring Your Challenges & Ask the Experts, Industrial Facilities /Inspection, Digital Historic Preservation, and University Lightning Round

The conference concludes with a plenary discussion on Experts View into the Future – What's Ahead for 3D Technologies, and Impact on Technology Investments.

The full conference program, with links to abstracts and speaker bios, is at www.spar3d.com/event/.

CALENDAR

- May 9-11, **15th Conference on Computer and Robot Vision (CRV 2018)**, Toronto, Ontario, Canada. For more information, visit http://www.computerrobotvision.org/.
- June 5-7, SPAR 3D Expo & Conference, Anaheim, California. For more information, visit www.spar3d.com/event.
- July 18-22, **COSPAR 2018**, Pasadena, California. For more information, visit http://www.cospar-assembly.org or http://cospar2018.org/.
- July 23-27, **URISA GIS Leadership Academy**, Salt Lake City, Utah. For more information, visit http://www.urisa.org/education-events/urisa-gis-leadership-academy/.
- August 19-23, **SPIE– Imaging Spectrometry XXII: Applications, Sensors, and Processing**, San Diego, California. For more information, visit http://spie.org/OPO/conferencedetails/imaging-spectrometry?SSO=1.
- September 19-20, **GIS in the Rockies**, Denver, Colorado. For more information, visit GISintheRockies.org.
- October 9-12, **GIS-Pro & CalGIS 2018: THE Conference for GIS Professionals**, Palm Springs, California. For more information, visit http://www.urisa.org/education-events/gis-pro-annual-conference/.

Productivity Analysis for Medium Format Mapping Cameras

Yuri Raizman
Phase One Industrial
Denmark

Introduction

Since 2000, development and use of digital photogrammetric cameras for aerial survey has gained significant momentum. Many different cameras and systems designed for aerial photogrammetry were developed and presented to the market. After 15 years of intensive development, only a few of these products are in wide use in today's mapping market. One of the prominent systems being provided is the medium format frame camera from Phase One Industrial.

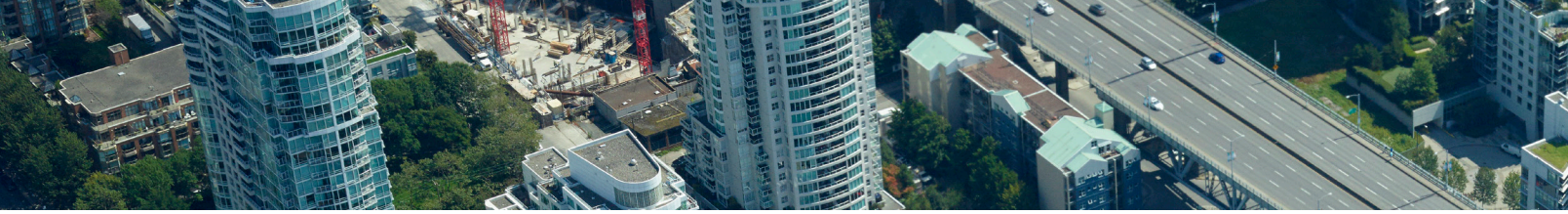
With the development of CCD and CMOS technology, medium format cameras have come a long way from 40-60 Mpix to 80-100 Mpix cameras. Additionally, high quality metric lenses with a wide range of focal lengths were developed and implemented. This enabled an effective utilization of medium format cameras in many different small and medium sized urban and rural mapping projects, corridor mapping, oblique projects, and monitoring of areal and linear infrastructure.

This article presents recent development in the approach to flight planning and aerial survey productivity analysis, firstly presented in Raizman (2012). The Raizman (2012) article referred only to large format cameras, whereas this article will compare large format cameras vs. medium format cameras, which are getting more and more popular in aerial survey. This approach is based on some pre-defined common characteristics of the required mapping products. It enables an equivalent comparison between cameras with different parameters – focal length, sensor form and size, and pixel size. Through this article we intend to demonstrate that for several types of urban mapping projects, medium format cameras and large format cameras have the same aerial survey productivity.

Image captured by Phase One iXU-RS1000, 70mm lens, Height 1,500ft, GSD 3cm.

Photogrammetric Engineering & Remote Sensing
Vol. 84, No. 5, May 2018, pp. 235–238.
0099-1112/18/235–238

doi: 10.14358/PERS.84.5.235



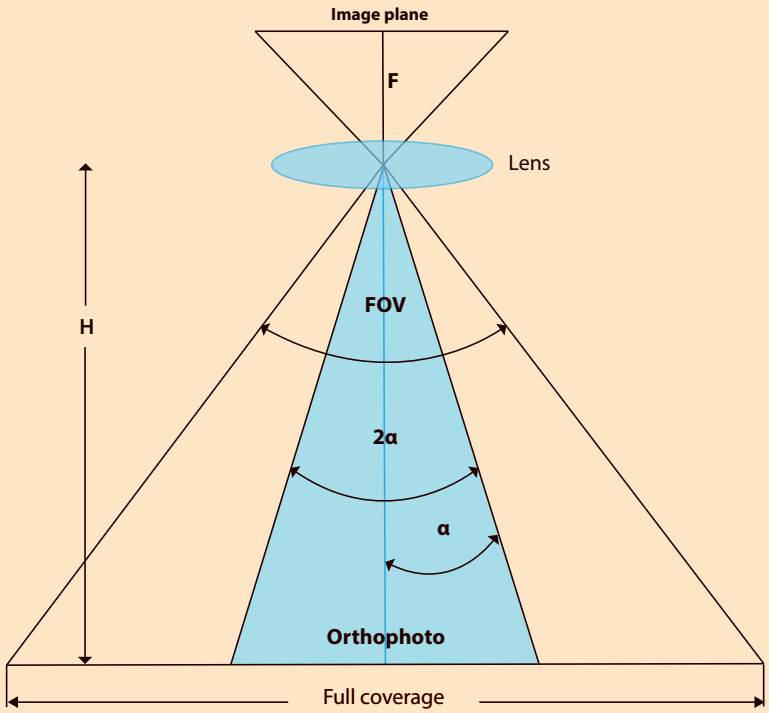
Common Denominator for Aerial Survey Cameras Comparison

There are two groups of aerial survey cameras – medium and large format metric cameras. There are also two main different types of mapping areas – urban and rural. There are three main photogrammetric products that are often required by the market – orthophoto, dense DSM (Digital Surface Model), and stereo mapping. We shall analyze the usage of these cameras for different applications.

One of the most popular products for urban area is a semi-true orthophoto. It features very narrow orthophoto angle (an effective angle, which is part of the Field of View used for orthophoto production and is equivalent to the required small building lean; see Figures 1 and 2) and very high level of visibility with minimizing hidden, shaded or obscured areas in the dense urban environment (Raizman, 2012). Figure 1 illustrates the central projection camera, FOV, orthophoto angle dedicated for orthophoto area on the images.

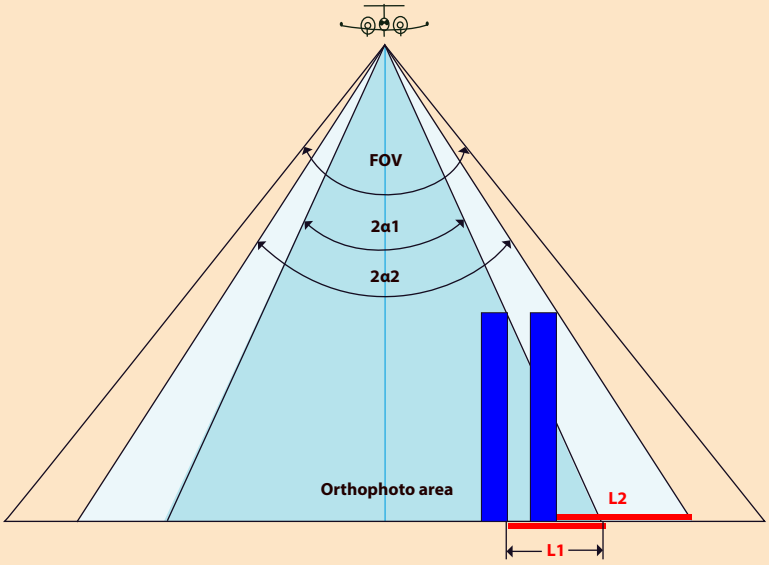
The concept of building lean and its importance for orthophoto is presented in Figure 2.

Ground resolution (or ground spacing distance, GSD) of 5 to 15 cm is commonly used for urban mapping. Orthophoto angles for orthophoto production in urban environment lie in the range of 14° to 25°, which corresponds to 12% to 22% of building lean. This predefined orthophoto angle (or building lean), GSD and minimal allowable side overlap are the three geometric parameters of aerial survey which enable a geometrically identical orthophoto (with the same building lean) from different aerial survey cameras. These three parameters are considered as a common denominator for a productivity comparison of different cameras of different types.



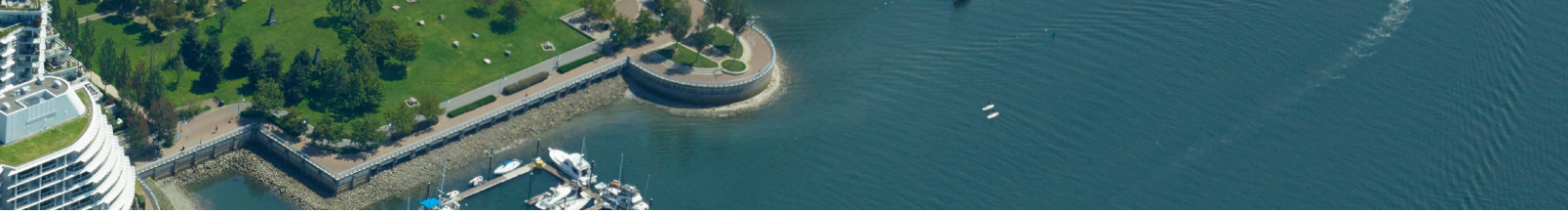
- F – Focal length;
- H – Flight altitude;
- FOV – Field of View, generally 27° - 110° for different aerial survey cameras;
- 2α – orthophoto angle;
- Tg(α) * 100% - Building lean.

Figure 1. Field of View and the Orthophoto Angle.



- 2α1, 2α2 – Permissible orthophoto angles;
- L1, L2 – Occlusion
- Tg(α)*100% - Building lean
- If 2α2 > 2α1 then L2 > L1

Figure 2. Field of View, Orthophoto Angle and Building Lean.



Productivity Comparison Between Medium and Large Format Cameras

Productivity comparison is commonly based on the following parameters: aerial survey productivity (image coverage per hour of flight), distance between flight lines, time required to fly Area of Interest (AOI) or number of flight lines per AOI. A more objective criterion, not depending on the ground speed of the plane and the shape of AOI, is the distance between flight lines. The following orthophoto geometrical parameters were used for calculations:

GSD	Orthophoto angle	Building lean	Ground Speed	Minimal side overlap
5 cm	14°	12%	120 knot	25%
8 cm	17°	15%	140 knot	25%
10 cm	20°	18%	160 knot	25%
15 cm	25°	22%	180 knot	25%

Based on the above assumptions, the following charts and tables present the productivity comparison for Phase One medium format cameras, Vexcel UltraCam Eagle and Hexagon DMC III large format cameras. Corresponding focal lengths of the cameras are presented in parenthesis.

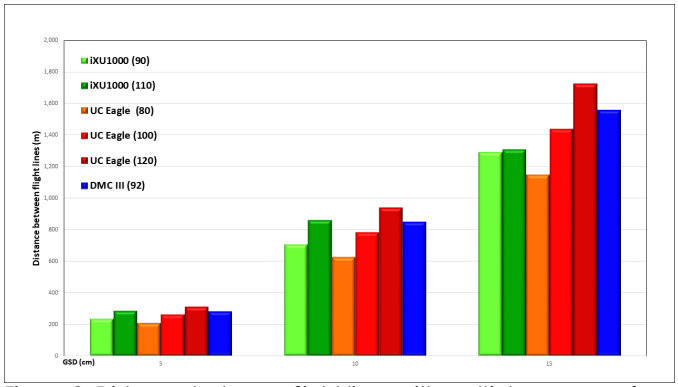


Figure 3. Distance between flight lines with multiple cameras from Phase One, UC Eagle and DMC III for orthophoto at 5 - 15 cm GSD.

Figure 3 demonstrates that with the requirement for orthophoto angle/building lean for urban orthophoto, medium and large format cameras provide similar distance between flight lines.

Figure 4 presents the time of flight needed to cover an area of 5 km by 5 km.

The same conclusion can be drawn from Figure 4. The requirement for orthophoto angle/building lean in urban environment equals the productivity of medium and large format cameras.

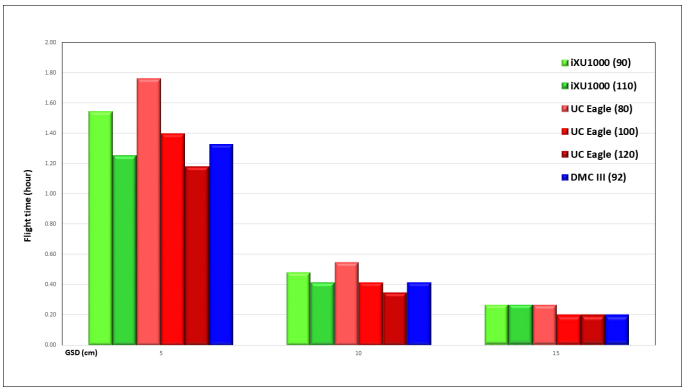


Figure 4. Flight time with Phase One, UC Eagle and DMC III for orthophoto at 5 - 15 cm ground sampling distance for an area of 5km x 5km.

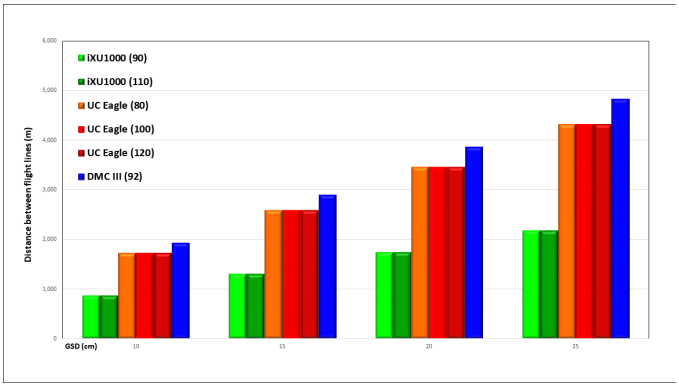
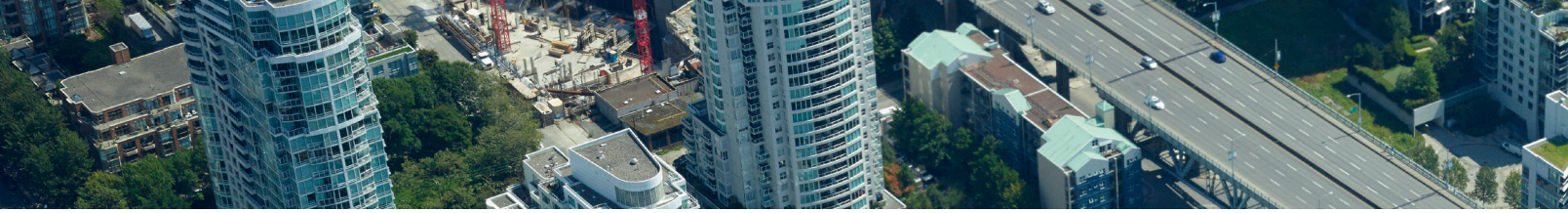


Figure 5. Distance between flight lines for rural area with 25% side overlap.

Figure 5 presents another situation common for other photogrammetric products: orthophoto for rural area, dense DSM or stereocompilation – flight without specific limitations on orthophoto angle with the minimal side overlap of 25% and with maximal use of the sensor (CCD/CMOS) area.

In this case, Phase One medium format cameras provide 50% of UC Eagle productivity and 45% of DMC III productivity, independently from the ground resolution. However, taking into consideration the relatively low purchase price of Phase One cameras, its utilization for medium size urban and rural mapping projects may be considered.

The wide range of exchangeable metric lenses with different focal lengths enables the use of Phase One cameras at different altitudes (Figure 6) with different flight platforms and for a variety of different purposes.



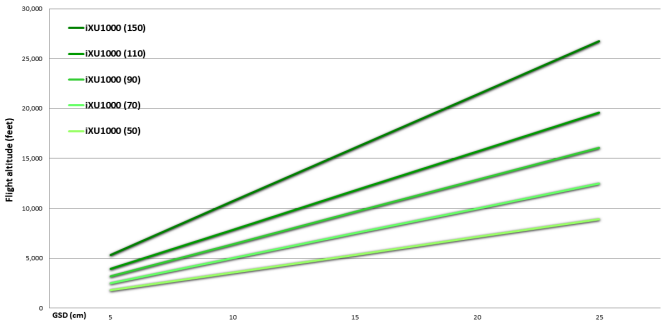


Figure 6. Flight altitudes with the wide range of Phase One metric lenses.

Conclusion

The last generation of Phase One medium format metric cameras with small pixel size (4.6 μ), large sensor area (100 Mpix), maximal frame-per-second (FPS) rate of 1.6 and exposure time of up to 1/2500 seconds, a set of metric lenses with different focal lengths (50, 70, 90, 110, 150 mm) and with relatively low price, provide an excellent alternative to large format cameras in many areas of aerial mapping and monitoring.

Additionally, these cameras are widely used for providing an oblique imagery and as a complementary camera for lidar systems. All these cameras, from oblique and from lidar systems, may be used as standalone cameras for mapping projects.

The very low weight (2 kg) and small size of the cameras enable their utilization with super-light planes, small helicopters, gyrocopters and UAVs, which can significantly reduce operational cost of mapping projects.

The Phase One cameras present an effective alternative to large format cameras for small and medium size urban and rural mapping projects, corridor mapping, oblique projects, and monitoring of areal and linear infrastructure.

Reference

Raizman, Y., 2012, Leaning Instead of Overlap – Flight Planning and Orthophotos, GIM International, June, pp. 35-38. (https://www.gim-international.com/content/article/flight-planning-and-orthophotos)

By Anita Simic Milas, PhD

Paradigm Shift in Education: Spatial Literacy

Remote sensing and geographic information system (GIS) technologies proliferate our mental capabilities of processing and combining spatial and textual information about various static or dynamic events and their interactions. An enormous amount of constantly accumulating remote sensing data evokes an urgent need for more remote sensing experts and for adding another dimension to the educational system in the U.S.A. and worldwide - Spatial literacy. The ability to process and analyze spatial data while addressing pressing environmental issues, is the fundamental form of literacy for experts in geoscience. The ultimate goal should be to promote the discipline of photogrammetry, remote sensing and geographic information systems (GIS) to enhance the understanding of the geospatial concepts, and to promote growing interest in the application of remote sensing by government, academia, and industry.

What are the Educational Gaps?

Integrating spatial literacy into the existing educational system is challenging. Remote sensing and GIS are subject to the existing obstacles of attracting students to the Science, Technology, Engineering and Mathematics (STEM) disciplines. High school students and their parents commonly perceive STEM fields as difficult and unrewarding.

The robust effort of the U.S. government to transfer remote sensing knowledge to K-12 students can be achieved only through the focused effort of experts, parents, teachers, graduate students and society. No simple solution answers the question "How to attract and keep students in Remote Sensing?" However, educational strategies that directly engage learners in the educational processes and techniques can increase the "STEM momentum".

The Cascade Education Model

While it is essential to engage high school students in STEM, children need to be even younger than high school age when they start learning about remote sensing. Thus, *parental expectation* plays a critical role in early learning. *K-12 teachers* further encourage students' curiosity and nurture their interest for STEM activity, like that of remote sensing. The knowledge transfer has to be envisioned in a cascading fashion where experts and university professors together with their university students educate both pre-service and in-service K-12 teachers who, in turn, emerge enthusiastic, knowledgeble and prepared to educate their students and other teachers. Some of the important components of the cascade model are:

- *Readiness of the teachers* to impart fundamentals of an academic discipline lays the foundation of success.
- *Peer learning* is a well-established concept where students learn from their peers while developing stronger self-esteem and better social skills through an active-learning.
- *Stimulating curiosity* about a student's spatial environment encourages the "what", "why", and "how" questions.
- *Interaction with industry and government experts* such as those from NASA/USGS/NOAA inspires students' interest and involvement in Earth and space science and provides them with an impressive depiction of possible achievements.
- *The 'learning by doing' concept* is the key component of active learning. To expose young students to problem-based learning at an early stage will improve their critical thinking and understanding.
- *Outreach and awareness about remote sensing,* through webinars and videos, inform and educate students, teachers, parents, young professionals and the public.

Recognizing some of the challenges and incorporating the components of the cascade model characterized was the focus of a recently accomplished project at Bowling Green State University (BGSU), in Ohio. The "*SPatial LITeracy - SPLIT Remote Sensing integrated research-educational approach to support surface water quality monitoring*" offered high

Photogrammetric Engineering & Remote Sensing
Vol. 84, No. 5, May 2018, pp. 239–240.
0099-1112/18/239–240

doi: 10.14358/PERS.84.5.239

school- and university-level students the opportunity to gain hands-on field remote sensing learning and research knowledge in an actual research scenario. This program was designed to help educate parents and K-12 teachers who could then contribute to creating a diverse and highly skilled future workforce in the field of remote sensing.

The project started with an exhibition that was open to the public called "SPatial LITeracy- *SPLIT through ART*" where students displayed over twenty visually-appealing remote sensing images. Each image was accompanied by its own story including information about remote sensing data and environmental issues. The goal was to spark the interest of students and the public in remote sensing through art. The event was advertised through various websites and the local newspaper (https://mobile.twitter.com/sentineltribune/status/913792707402878977/video/1).

Soon after the exhibition, the teaming effort between Mr. Roger Blevins (Huron High School, Ohio), Dr. Kristin Arend (Ohio Department of Natural Resources, Division of Wildlife), and Dr. Anita Simic Milas (BGSU), organized a series of field campaigns and educational events at the Old Woman Creek (OWC) National Estuarine Research Reserve, Huron High School and BGSU, where students and teachers learned how to acquire and process spectral information using the field spectroradiometer and UAV/drone. They derived spectral indices related to water quality using real data to map and correlate chlorophyll-a and total phosphorous in OWC. Knowledge was first passed from Dr. Simic Milas' graduate to undergraduate students and then, in separate sessions, to high school students and teachers.

As part of the project, a one-day workshop was organized at BGSU for the students and teachers of Huron High School and other local high schools. Mr. Roger Tokars and Mr. Rigoberto Roche of NASA Glenn Research Center, and Dr. Andrea Vander Woude from NOAA shared their knowledge about "*Monitoring Harmful Algal Bloom in Lake Erie*". Dr. Robert Vincent, Professor Emeritus of BGSU, presented his perspective on remote sensing research from his experience while working in academia and industry.

An experts-panel discussion, moderated by Dr. Simic Milas, focused on how to attract students into the field of remote sensing and how to keep them interested throughout high school and university. The panel comprised of an Associate Professor of Education, two experts in the remote sensing field, a high school science teacher with two of his students, and undergraduate and graduate students. The discussion underscored the importance of three factors for learning success: (1) Students must be provided with hands-on training and experience-designed projects in educational streams nowadays; (2) The educational system at universities should favor knowledge and interest over grade-based achievements; (3) Teachers should be additionally educated and more effective in passing remote sensing knowledge and enthusiasm about the field to their students.

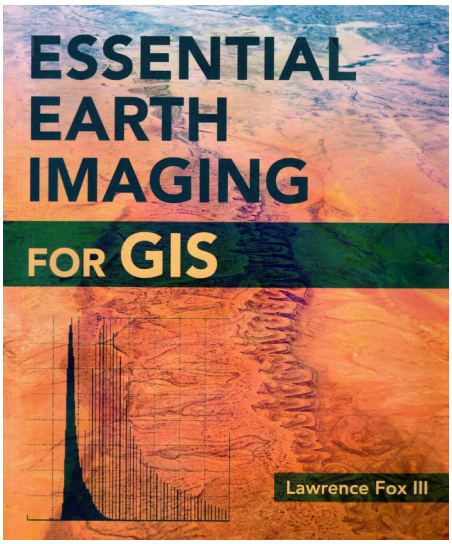
The workshop concluded with another component of the cascading education model, a hands-on session where high school students and teachers had a chance to learn the basics of image manipulation and visualization. The outreach component was enriched by the YouTube videos and webinars created by the university and high school students (https://www.youtube.com/channel/UCkmhoMQihvhRb1I8DpQ2wNw).

All components of the SPLIT Remote Sensing project promoted each other. While the SPLIT exhibition event, websites and newspaper articles attracted teachers and students to the program, the field campaigns and hands-on tutorials helped them to understand remote sensing basic concepts. An active-learning and team-building atmosphere were a driving force behind the peer learning between the students. The interaction between the students and NASA experts motivated students to participate in the panel discussion. The YouTube videos and webinars have secured ongoing engagement of teachers and students in the field of remote sensing after the event. Overall, the SPLIT Remote Sensing project could be described as an effective cascade education model that helps instantiate spatial literacy of students, teachers and parents. Moreover, the SPLIT Remote Sensing educational approach empowers the geospatial community by attracting more young people who would take the drive towards more rapid development of the disciplines of photogrammetry, remote sensing, GIS, and other supporting geospatial technologies. Thus, the role of the remote sensing societies, such as the American Society for Photogrammetry and Remote Sensing (ASPRS) and consortiums such as America View, in promoting the same or similar educational models is critical.

Funding for this activity was provided by AmericaView through a grant from the U.S. Geological Survey, awarded in December 2013.

Author

Dr. Anita Simic Milas is an Assistant Professor in the School of Earth, Environment and Society (SEES) at BGSU, Ohio. She is also a director of SPLIT Remote Sensing® (http://splitremotesensing.com/).



Essential Earth Imaging for GIS

Lawrence Fox III

Esri Press, 2015. ISBN: 9781589483453. Trim: 7.5in X 9 in.
Format: Trade paper. Price: $59.99. Pages: 128

Reviewed by Connie Li Krampf, CP, CMS and MSCS, Chief Photogrammetrist, Timmons Group, Raleigh, North Carolina.

In the GIS world, imagery has been used for geoprocessing for decades. Imagery has always been a very important component of Geospatial applications. With continually advancing imagery technology, it becomes even more so. GIS software provides various functionalities and tools for processing, displaying, and interpreting imagery and extracting the features from it. Understanding how the software works to process the imagery becomes a normal, required task for geospatial professionals to do the work efficiently. The book *Essential Earth Imaging for GIS* written by Lawrence Fox III provides rich information and guidelines in GIS imagery technology not only for geospatial professionals but also for the college student, as a reference for introductory GIS courses that include multispectral imagery display and analysis.

Essential Earth Imaging for GIS provides a basic education in imaging technology and management, promoting the effective use of imaging tools in GIS software. This book includes concepts and methods of image formation and manipulation that enable the user to efficiently and effectively display, co-register, enhance, interpret and delimit features from earth imagery.

The book opens with a short but thorough introduction that orients the book's audience, the goals of the book, and explains how and why the book is organized as it is. The outline of each chapter is provided. It is a slim book with eight chapters, references, and index.

Chapter 1, "Overview of Imaging GIS," briefly describes the earth imagery history while it summaries the types of imagery used in GIS based on the sensor systems that include satellites, aircraft, and unmanned aerial systems (UAS). The chapter identifies the structure of the two–dimensional digital imagery in which the value of each cell or pixel represents the brightness of imagery. With this important concept, the author points out that "*the interpreter can use the geographic pattern of relative brightness values to help correctly interpret Earth features*…." Right after this important perspective, another fundamental but crucial concept of color imagery is introduced: "...all colors can be formed from various shades of three additive primary colors: red, green and blue. Every color image is actually three images superimposed in various shades of those three colors. In the same chapter, the author also discusses the Three-dimensional data that is used for three-dimensional visualization of terrain in GIS software.

Chapter 2, "The Physical Basis and General Methods of Remote Sensing," presents the most important learning point of the book. The author uses straightforward language to explain the principles of electromagnet (EM) radiation, the engine of image formation in virtually all earth imaging systems. The author covers the complex science of imaging systems, the capabilities, and limitations of various remote sensing methods, the aerial and spaceborne platforms and how their characteristics influence the attributes of the imagery collected from them with simple diagrams, pictures, and detailed explanations. This is one of the longest chapters in the book. The author invests much effort in demonstrating the theories and the advanced level science behind the remote sensing technology, which the reader should appreciate.

Chapter 3 "Effects of the Atmosphere on Image Quality" is equally important for the user to understand. In this chapter, the user will learn how atmosphere and cloud cover affects the electromagnetic (EM) radiation detected by remote sensing systems. The author clearly presents for the reader very important concepts and facts related to the topic by using numbers, diagrams, and pictures in addition to the detailed explanations. One example of the easy to understand statements but very important concepts in the chapter is "When mapping surface features with imaging GIS, the reflected or emitted radiation is the *signal, and the atmospheric contribution is the noise."*

Photogrammetric Engineering & Remote Sensing
Vol. 84, No. 5, May 2018, pp. 241–242.
0099-1112/18/241–242

doi: 10.14358/PERS.84.5.241

Chapter 4 "Creating Two-dimensional Images with Sensors" educates the user to understand how sensors work to effectively use the two-dimensional images generated by remote sensors. The chapter covers the details of instruments that generate two-dimensional images, including cameras, multispectral sensors, and imaging radar. In this chapter, the most fascinating section would be "General image attributes: The Four Rs." The author explains the imagery attributes in great detail because "*an understanding of four attributes of images allows practitioners to evaluate different types of remote sensing images for many applications regardless of the sensor technology used to produce them*." The four Rs are the characteristics of the imagery resolution. They all exist in our daily practice when we, as geospatial professionals, produce imagery or use it in our GIS analysis. Accordingly, understanding these four Rs and applying this knowledge in our work will help us do our work efficiently and deliver the best results possible to our customers.

Chapter 5, "Displaying Digital Images with GIS Software" opens with a talk about human vision and the engineer's ability to mimic human color perception to make color photography, television, and remote sensing imagery. In the section "True-color Images," the author renders the history of "the tristimulus theory of color vision…key to the development of color imaging.." and reviews the theory with colorful diagrams. He provides examples of the application in image display and printing technology. In the section "Assigning spectral bands to colors," the author leads us to an even broader knowledge base by teaching us to understand how spectral bands are assigned to colors and how to learn to assign them in effective ways to produce false colors to help us to gather more information in remote processing. The section "How Software Controls Contrast and Brightness of Color Displays" discusses the details and mathematics of histograms and presenting image brightness. The section "Stretching the Histogram of a Single-band Image to Enhance Contrast and Brightness" in GIS software provides important guidelines for practitioners to understand how to use the tools to achieve the best imagery display when working with the imagery. The last section of the chapter introduces "Pseudo Color Images" that should not be confused with false-color images, as pointed out by the author.

Chapter 6, "Generating Three-dimensional Data with Photogrammetric Measurements and Active Sensors", introduces the audience to another important topic regarding earth imagery. The author overviews the technologies that generate three-dimensional data including photogrammetry, lidar, and interferometric radar technology. The section "Obtaining Vertical and Horizontal Positions from Aerial Photographs" includes subsections "Geometry of a Single Aerial Photography," "Geometry of an Overlapping Pair of Aerial Photographs," "Digital Surface Models and Orthophotos," and "Incorporating Machine Vision into Photogrammetry" which cover important concepts, technologies, and methodologies of Photogrammetry. The section "Obtaining Vertical and Horizontal Positions from Lidar Sensors" presents how the lidar systems work. The last section "Obtaining Vertical and Horizontal Positions from Interferometric Radar Sensors" briefly but precisely describes how interferometric radar works to produce three-dimension data.

Chapter 7, "Image Processing," includes a typical workflow illustrating image processing procedures that are normally performed by imager providers, image analysts and GIS software users. The chapter provides the detailed technical knowledge and guidelines for Imager Restoration," "Image Rectification," and "Imager Enhancement." It also discusses challenges and technology in the "Conversion to Radiance" and "Atmospheric Correction" in imaging processing. Toward the end of the chapter, the author briefly prescripts "Image Processing in the Cloud" that has been increasingly popular in the geospatial industry with advancing internet technology. In the last section of this chapter, "Typical Workflow for Image Processing," a workflow chart demonstrates the procedures used in the image processing while the summary of each step in the flowchart is clearly stated.

Chapter 8, the final chapter, "Extracting Information from Images" provides very important and useful guidelines for GIS professionals to perform the tasks of imagery interpretation and delineation using GIS software. The user will also learn the advantages and disadvantages of automated image classification methods and how to evaluate maps generated using these methods. In the same chapter, the author points toward the future developments of the technology.

The book is well organized so that the user can follow along easily. The author explains the details of each technical topic with straightforward language and professionally. The book provides many useful examples, pictures, diagram and tables, which makes the book easy and fun to read. Moreover, the user can actually get hands on the GIS software and do the excises with real data. The exercises are thoughtfully designed and the data carefully selected. The step by step, easy follow instructions are included in the exercise materials. These exercises help the user to understand the book in the context of the professional industry.

The author provides detailed information in the Introduction of the book about how and where to download the 180-day trial of ArcGIS software and associated exercise instructions and data. The book is well written and provides very useful guidelines for GIS professionals, as well as great resources for students who are pursuing a geospatial career. Even if one is already familiar with GIS software, he/she can still derive great benefits from studying the book to get a better understanding of earth imagery, which should benefit them in their daily work. My only suggestion might be that it should be more convenient for the user if the Exercise Instructions were made part of the book as an appendix. Regardless, it is a great book and strongly recommended for geospatial professionals who are working with Earth imagery and students who want to understand Earth imaging in GIS processing and analysis.

& GRIDS & DATUMS

THE REPUBLIC OF POLAND

BY Clifford J. Mugnier, CP, CMS, FASPRS

The Grids & Datums column has completed an exploration of every country on the Earth. For those who did not get to enjoy this world tour the first time, *PE&RS* is reprinting prior articles from the column. This month's article on The Republic of Poland was originally printed in 2000 but contains updates to their coordinate system since then.

In the 10th century Poland was a Slavic duchy. The crown became elective after 1572, and various wars caused much loss of territory. Napoleon partly reestablished the kingdom (1807-15), which was later to become closely aligned with Russia. Poland was declared a republic in 1918, but it wasn't until after nearly another century of being overrun and controlled by others, that the new constitution was dated 16 October 1997.

The north and central regions are essentially flat and characterized by morainic topography. This lack of natural barriers on the North European Plain was a major reason for so many invasions of Poland through-out history. The southern boundary is mountainous, with the highest peak being Rysy at 2499 m (8197 ft.); the lowest point in Poland is Raczki Elblaskie at –2 m.

Early mapping of Poland was instituted by the Prussians for the western half of the present country, and ap-proximately 17% of the southeast was mapped by the Austro-Hungarian Empire. The remainder of Poland was surveyed and mapped by czarist Russia. The date of this early mapping activity goes back to 1816. The early Prussian Landesaufnahme characteristically used the Cassini-Soldner projection in its spherical form that was based on equivalent (Gaussian) spheres referenced to the Bohnenberger ellipsoid and the Zach ellipsoids, and later the Bessel 1841 ellipsoid. [See also the Republic of Hungary (*PE&RS* April'99) and the Czech Republic (*PE&RS* Jan '00).] The Prussians and the Austrians introduced the concept of the cadaster, or system of surveys and land registration for ownership and taxation. The Austro-Hungarian surveyors had similar preferences for ellipsoids, but the Russians were a different story.

The tsarist Russians performed surveys and topographic mapping of Poland in the 19th and early 20th centuries, but these works were for military purposes only. They did nothing with respect to individual land ownership registration,

and they preferred the sazhen for their unit of measurement. In the years between the two world wars, this source material was responsible for some very strange-looking contour maps of Poland when the unit of measurement was changed from sazhens to meters where 1 sazhen = 2.134 meters. (The only time I see similar strange values for contours is when I grade some of my sophomores' campus topo maps). The Russians preferred the Walbeck 1819 ellipsoid where a = 6,376,896 meters and the reciprocal of flattening, $(^1/_f)$ = 302.78. Some of these old maps also referred longitudes to Ferro in the Canary Islands; a practice dropped after WW II.

New geodetic triangulation started after the founding of the re-public, and the origin of the Polish National Datum of 1925 (PND 1925) is at station Borowa Gora (gora is Polish for mountain) where:(Φ_o) = 52° 28′ 32.85″ North, and (Λ_o) = 21° 02′ 12.12" East of Greenwich. The ellipsoid

Photogrammetric Engineering & Remote Sensing
Vol. 84, No. 5, May 2018, pp. 243–245.
0099-1112/18/243–245

doi: 10.14358/PERS.84.5.243

of reference is the Bessel 1841 where a = 6,377,397.155 and $^1/_f$ = 299.1528128. Instruments used by the Polish military included theodolites manufactured by Bamberg, Fennel, Wild-Heerbrugg, and Aerogeopribor. Over 120 triangles were observed from 1927 to 1935, with the average angular error of figure not exceeding 0.56 arc seconds. (Post WW II observations with T-3 theodolites yielded errors not exceeding 0.46 arc seconds). Baselines were measured at Grodno, Kobryn, Warsaw, Lomża, Luniniec, Mir, and Braslaw. Laplace stations (astro shots) were observed at Varsovic (initial point Borowa Góra), Borkowo, Kopciowka, and Skopowka.

The 1:25,000, 1:100,000, and 1:300,000 maps produced by the Wojskowy Instytut Geograficzny (WIG) or Military Geographical Institute, were products of carto-graphic datum shifts (scissors and paste) to PND 1925 and are cast on the Polish Stereographic Grid. The old Polish Stereographic Grid is based on the mathematical model by Rousilhe, who was the Hydrographer of the French Navy. All ellipsoidal oblique stereographic projections developed and used worldwide before WWI are based on Rousilhe's work that was originally published in *Annals Hydrographique*. The development of the projection for nearby Romania (STEREO 70) was done by the Bulgarian geodesist, Hristow in the late 1930's. The PND 1925 WIG Military Stereo-graphic Latitude of Origin (ϕ_o) = 52° 00′ N, Central Meridian (λ_o) = 22° 00′ E, the Scale Factor at Origin (m_o) = 1.0, the False Easting = 600 km, and the False Northing = 500 km. The PND 1925 Grids developed for the Cadastre around the same time were cast on the Gauss-Krüger Transverse Mercator where the Scale Factor at Origin (m_o) = 1.0, the False Easting = 90 km, the False Northing = *minus* 5,700 km *at the equator*, and the Central Meridians (λ_o) = 15°, 17°, 19°, and 21° East of Greenwich.

During WWII, the Generalstab des Heerres, Reichsamt für Landesaufnahme (German Army) produced topographic maps of Poland cast on the Deutches Herres Gitter (DHG) Grid, which is identical to the UTM except for the Scale Factor at Origin (m_o) = 1.0. Of course, the Datum used was the PND 1925, as was the equivalent treatment of Poland by the USSR with the Russia Belts TM that had the same defining parameters as the DHG except for the Datum and ellipsoid. The Russian coverage during the war had *double* Grids in Poland that exhibited unresolved horizontal Datum discrepan-

cies ranging from 160 to 250 meters. After the war, the USSR converted their *military* topographic coverage of the Warsaw Pact countries to the System 42 Datum that has its origin at Pulkovo Observatory and is referenced to the Krassovsky ellipsoid. In the Republic of Poland, their preferred terminology of that Datum is "Polish National 1942" or "PN 42."

Large scale topographic maps of Poland published in the latter part of the 20th century are on the "UKŁAD 65 System," the parameters of which have been a closely held secret. In the past few years, little information has dribbled out of Poland on these "Strefa" or Zones. In February of 2000, Wojtek Hanik sent the de-classified parameters to me! There are five Strefa comprising the UKŁAD 65 System, four are based on the "Quasi-Stereographic" Grid (Rousilhe Oblique Stereographic), and the fifth is a Gauss-Krüger Transverse Mercator Grid. Strefa 1 covers the following provinces: Biala Podlaska, Eastern Bielsko, Chełm, Kielce, Kraków, Krosno, Lódź, Lublin, Nowy Sącz, Piotrków, Premysl, Radom, Rzeszów, Sieradz, Tarnobrzeg, Tarnów, and Zamość. The UKŁAD 65 Strefa 1 Quasi-Stereographic Grid Latitude of Origin (ϕ_o) = 50° 37′ 30″ N, Central Meridian (λ_o) = 21° 05′ 00″ E, the Scale Factor at Origin (m_o) = 0.9998, the False Easting = 4,637 km, and the False Northing = 5,467 km. Strefa 2 covers the following provinces: Bialystok, Ciechanów, Lomża, Olsztyn, Ostrolęka, Plock, Siedlce, Skierniewice, Suwałki, and Warszawa. The UKŁAD 65 Strefa 2 Quasi-Stereographic Grid Latitude of Origin (ϕ_o) = 53° 00′ 07″N, Central Meridian (λ_o) = 21° 30′ 10″ E, the Scale Factor at Origin (m_o) = 0.9998, the False Easting = 4,603 km, and the False Northing = 5,806 km. Strefa 3 covers the following provinces: Bydgoszcz, Elbąg, Gdańsk, Koszalin, Słupsk, Szczecin, Toruń, and Włocławek. The UKŁAD 65 Strefa 3 Quasi-Stereographic Grid Latitude of Origin (ϕ_o) = 53° 35′ 00″ N, Central Meridian (λ_o) = 17° 00′ 30″ E, the Scale Factor at Origin (m_o) = 0.9998, the False Easting = 3,703 km, and the False Northing = 5,627 km. Strefa 4 covers the following provinces: Gorzów, Jelenia Gora, Kalisz, Konin, Legnica, Leszno, Opole, Pila, Poznań, Wałbrzych, Wrocław, and Zielona Góra. The UKŁAD 65 Strefa 4 Quasi-Stereographic Grid Latitude of Origin (ϕ_o) = 51° 40′ 15″ N, Central Meridian (λ_o) = 16° 40′ 20″ E, the Scale Factor at Origin (m_o) = 0.9998, the False Easting = 3,703 km, and the False Northing = 5,627 km. Strefa 5 covers the following provinces: Western Bielsko, Częstochowa, and Katowice. The UKŁAD 65 Strefa 5 Gauss-Krüger Transverse Mercator Grid Central Meridian (λ_o) = 18° 57′ 30″ E, the Scale Factor at Origin (m_o) = 0.999983, the False Easting = 237 km, and the False Northing = *minus* 4,700 km.

For small scale mapping, the GUGiK 80 Quasi-Stereographic (Rousilhe) projection is used where the Latitude of Origin (ϕ_o) = 52° 12′ N (approx.), Central Meridian (λ_o) = 19° 10′ E. The scale factor at a point is designed to be equal to unity at a distance of 215 km from the projection origin. These mysterious parameters, of which some are still held and some are now public, reflect the history of the nation. Those countries that have a long and recent history of war, occupation, or blood spilled at borders will be particularly sensitive about releasing Grid and/or Datum relation parameters. The release of some of this previously secret data may be an indication of the Republic's confidence in the future.

Update

"The ETRS89 was introduced in Poland technically by the GNSS technique in the last years of the 20th century and by law in 2000. On 2 June 2008, the Head Office of Geodesy and Cartography in Poland (GUGiK) commenced operating the multifunctional precise satellite positioning system named ASG-EUPOS. The ASG-EUPOS network defines the European Terrestrial Reference System ETRS89 in Poland. A close connection between the ASGEUPOS stations and 18 Polish EUREF Permanent Network (EPN) stations controls the realization of the ETRS89 on Polish territory. In 2010-2011 GUGiK integrated the ASG-EUPOS with the existing geodetic networks (horizontal and vertical) using GNSS and spirit levelling. Those actions resulted in developing and then legal introduction in 2012 new technical standards: to the National Spatial Reference System (PSOP) and to establish and maintain the geodetic (horizontal and vertical), gravity and magnetic control in the country. Thus, the geodetic, gravimetric and magnetic system in Poland has been associated with the European one (previous and current). This allowed for the next step of networks integration in Poland, namely, in 2013 started integration of national geodetic control with gravimetric control. Modern geodetic, gravimetric and magnetic networks in Poland are to be fully consistent with the European system. In 2011, following the initiative by the Section of Geodetic Networks and the Section of Earths' Dynamics of the Committee on Geodesy of the Polish Academy of Sciences, a new research network "Polish Research Network for Global Geodetic Observing System" (acronym GGOS-PL) has been established" *(Jarosław Bosy, and Jan Kryński, POLISH NATIONAL REPORT ON GEODESY 2011– 2014, Vol. 64, Warsaw, 2015).*

This column was previously published in *PE&RS*.

ASPRS MEMBERSHIP

ASPRS would like to welcome the following new members!

At Large
Mengge Chen
Zhuo Chen
Jamey Gray
Yue Gu
YING Li
Ming Liu
Weiya Ye

Columbia River
Ratnanjali Adhar
Ian Campbell Davis

Eastern Great Lakes
Kevin Corbin
Ross Mellgren
Frank Tokar, Jr.

Florida
John Blair Northrop

Heartland
Brandon Kyle Conlon
Trey Olen Lee

Mid-South
Bradley G. Bishop

Pacific Southwest
Jason Owens
Amy Work
Vince Zaragoza

Potomac
Karen E. Stone, PLS
Chelsea White
Yadong Xu

Rocky Mountain
Andrew Archer
Prof. Michael Lefsky
Edward J. Norero, Jr.
Isidore Uwalaka

Western Great Lakes
Sharon White

FOR MORE INFORMATION ON ASPRS MEMBERSHIP, VISIT
HTTP://WWW.ASPRS.ORG/JOIN-NOW

Your path to success in the geospatial community

Who at ASPRS Do I Contact to…

425 Barlow Place, Suite 210, Bethesda, MD 20814
301-493-0290, 301-493-0208 (fax),
www.asprs.org

Membership/*PE&RS* Subscription/Conferences
Priscilla Weeks — pweeks@asprs.org, x 109

Advertising/Exhibit Sales
Bill Spilman — bill@innovativemediasolutions.com

Peer-Review Article Submission
Alper Yilmaz — PERSeditor@asprs.org

Highlight Article Submission
Jie Shan — jshan@ecn.purdue.edu

Calendar
calendar@asprs.org

Editorial
Best of "ISPRS Hannover Workshop 2017"

Guest Editors Christian Heipke, Karsten Jacobsen, and Franz Rottensteiner, Uwe Stilla, Michael Ying Yang, Jan Skaloud, Ismael Colomina, and Michael Cramer

Sensor calibration, image orientation, object extraction and scene understanding from images and image sequences are important research topics in Photogrammetry, Remote Sensing, Computer Vision and Geoinformation Science, the areas of interest of the International Society for Photogrammetry and Remote Sensing (ISPRS). Within these areas, both geometry and semantics play an important role, and high quality results require appropriate handling of all these aspects. While individual algorithms differ according to the imaging geometry and the employed sensors and platforms, all mentioned aspects need to be integrated in a suitable workflow to solve most real-world problems.

This observation led to the organization of a common event for a number of well-established scientific meetings under the roof of the ISPRS Hannover Workshop, held in Hannover, Germany from June 6 – 9, 2017. These meetings were

- HRIGI - High-Resolution Earth Imaging for Geospatial Information, which has been held in Hannover every two years since the middle of the 1990's,
- CMRT - City Models, Roads and Traffic, a workshop dealing with automatic object extraction in urban environments with a first edition in 2005,
- ISA - Image Sequence Analysis, a relatively new workshop focussing on images sequences,
- EuroCOW - European Calibration and Orientation Workshop, looking specifically at sensors, calibration and orientation, which had previously been held in Barcelona, Spain for many years.

While HRIGI and EuroCOW are more on the geometric side, CMRT and ISA have a legacy in automatic object reconstruction and trajectory computation. The aim of the common event was to seek, exploit and deepen the synergies between geometry, semantics and sensor modelling, and to give the different scientific communities the possibility to discuss with, and to learn from, each other. The joint event was supported by 12 working groups from four of the five ISPRS Technical Commissions[1] and addressed experts from research, government, and private industry. It consisted of high quality papers, and provided an international forum for discussion of leading research and technological developments, as well as applications in the field.

Following the workshop, authors whose contributions were accepted after a full-paper double blind review were invited to revise and extend their papers in the light of the discussions at the workshop and to submit them to a special issue of Photogrammetric Engineering & Remote Sensing. Sixteen papers were submitted and after another round of scientific reviews, eleven of them were finally accepted for publication in this special issue. This large number constitutes a major success and demonstrates both, the relevance of the addressed topics and the high quality of the manuscripts; it has also led to the fact that the special issue had to be distributed to two volumes.

The first volume contains papers related to the classification of images (three papers) and point clouds (two papers) and to change detection (one paper). The first two papers of the second volume deal with sensor design and calibration, the following two with point cloud segmentation and the last two with the modelling of specific topographic objects (buildings in this case).

The first paper, authored by Vogt et al. and entitled Unsupervised source selection for domain adaptation deals with transfer learning, i.e. the question to which extent training data from one geographic area or epoch (called source) can be employed to classify data of another area or epoch (target) even if the features in the target image follow a slightly different distribution. More specifically, the best among many available sources for a specific classification problem is determined based on similarity measurements between the marginal distributions of the features in the source and various target domains.

The second paper, Multitemporal classification under label noise based on outdated maps by Maas et al. is devoted to the problem arising from incorrect training data. While for map updating abundant training data are available in the form of the (outdated) map itself, some for the training data are incorrect and result in wrong classification results. The authors develop a new noise tolerant classification method that can also consider the outdated map as prior information and show that it helps to distinguish between real changes over time and false detections caused by misclassification.

The paper by Drees and Roscher, Archetypal analysis for sparse representation-based hyperspectral sub-pixel quantification,

[1] https://www.isprs-ann-photogramm-remote-sens-spatial-inf-sci.net/IV-1-W1/1/2017/isprs-annals-IV-1-W1-1-2017.pdf, doi.org/10.5194/isprs-annals-IV-1-W1-1-2017, 2017 for details

Photogrammetric Engineering & Remote Sensing
Vol. 84, No. 5, May 2018, pp. 247–248.
0099-1112/18/247–248

doi: 10.14358/PERS.84.5.247

suggests a new classification method for hyper-spectral image data. Typically these data have a rather coarse geometrical resolution resulting in mixed pixels (pixels containing more than one spectral class). The authors develop a new constrained sparse representation of the data, where each pixel with unknown surface characteristics is expressed by a weighted linear combination of elementary spectra with known land cover class. They then determine the elementary spectra from image reference data using archetypal analysis combined with a Reversible Jump Markov Chain Monte Carlo method.

The next group of two papers on 3D point classification starts with Classification of aerial photogrammetric 3D point clouds by Becker et al. The authors present a new method to classify 3D point clouds derived from aerial imagery which exploits both geometric and colour information. They show that incorporating colour yields a significant increase in accuracy; the approach can also be used to derive high accuracy digital terrain models from digital surface models.

The contribution by Hackel et al. entitled Large-scale supervised learning for 3D point cloud labelling: SEMANTIC3D.NET suggests a new benchmark data set for the classification of 3D point clouds. Inspired by the recent success of deep learning and Convolutional Neural Networks (CNNs), attributed to the large number of employed training data, the authors hope to boost 3D point classification in a similar way by providing a total of four billion manually labelled points for investigation by the scientific community. They also describe some initial work that under pins the expectations that CNNs might also lead to very competitive results in this area.

The authors of the last paper of the first volume, Yang et al., work on the problem of building change detection. In 4D Change detection based on persistent scatterer interferometry - A case study of monitoring building changes they suggest to track persistent scatters over time and to use them as indicators for new and demolished buildings, respectively, in an automatic statistics-based scheme. The new approach is successfully evaluated based on simulations and on TerraSAR-X images.

The second volume stars with two papers on sensors. In the first contribution, On a novel 360° panoramic stereo mobile mapping system, Blaser et al. present a new mobile mapping system equipped with different panoramic cameras which achieves a full 360° multi-stereo coverage. The authors report on system calibration and operational tests which yielded an accuracy in the cm to dm range for both, relative and absolute measurements.

The next paper, authored by Voges et al., deals with a particularly important and often overlooked aspect for sensor system calibration, namely time synchronisation. Using an example from robotics the authors show how time offsets between different parts of the sensor system can be retrieved for SLAM (simultaneous localisation and mapping) observations.

The next two papers are concerned with the non-semantic segmentation of point clouds from different sources. The contribution A voxel- and graph-based strategy for segmenting 3D buildings scenes using perceptual grouping laws: comparison and evaluation by Xu et al. presents two different segmentation methods using voxel and supervoxel data structures, respectively, by help of perceptual grouping. In experiments using both laser scanning and photogrammetric point clouds the authors could demonstrate high quality results also for complex scenes and nonplanar object surfaces.

In their paper Range-image: Incorporating sensor topology for lidar point clouds processing, Biasutti et al. take a different view on LiDAR point cloud processing. Rather than working in 3D they project the 3D points into 2D space, arguing that in this way the large amount of successful work on disocclusion from images can be made use of. Based on these images a semi-automatic segmentation procedure based on depth histograms is presented, and detected occluded areas are reconstructed using a variational image inpainting technique.

The last two papers of this special issue tackle the problem of object modelling. First, in Geometric reasoning with uncertain polygonal faces, Meidow and Förstner discuss different strategies which can help to strike a balance between too unspecific and too restrictive models. They then suggest to model and to instantiate buildings as arbitrarily shaped polyhedra and to recognize man-made structures in a subsequent stage by geometric reasoning; examples are given to illustrate their method.

We hope that the reader will enjoy this variety to papers ranging from sensor design to semantic image and point cloud processing, and from novel scientific techniques to the investigation of data acquisition and processing systems. We would like to sincerely thank everybody involved in the preparation of this special issue. First of all and foremost, we are very grateful to Alper Yilmaz, Editor-in-Chief of PE&RS, to have offered to us the possibility to publish refined versions of the workshop manuscripts in his journal, and for all the freedom we could enjoy when preparing the special issue. We are very grateful to the authors of this special issue for making available their excellent papers, and for keeping a tough timeline. We also wish to wholeheartedly thank the reviewers of both, the workshop and the journal papers, who have tremendously contributed to improve the submitted manuscripts. We wish you, the readers, an informative and enjoyable reading and hope that we could reach the level of scientific excellence you expect from this journal.

The Guest Editors: Christian Heipke, Karsten Jacobsen, and Franz Rottensteiner (Hannover), Uwe Stilla (München), Michael Ying Yang (Enschede), Jan Skaloud (Lausanne), Ismael Colomina (Casteldefels) and Michael Cramer (Stuttgart)

Unsupervised Source Selection for Domain Adaptation

Karsten Vogt, Andreas Paul, Jörn Ostermann, Franz Rottensteiner, and Christian Heipke

Abstract

The creation of training sets for supervised machine learning often incurs unsustainable manual costs. Transfer learning (TL) techniques have been proposed as a way to solve this issue by adapting training data from different, but related (source) datasets to the test (target) dataset. A problem in TL is how to quantify the relatedness of a source quickly and robustly. In this work, we present a fast domain similarity measure that captures the relatedness between datasets purely based on unlabeled data. Our method transfers knowledge from multiple sources by generating a weighted combination of domains. We show for multiple datasets that learning on such sources achieves an average overall accuracy closer than 2.5 percent to the results of the target classifier for semantic segmentation tasks. We further apply our method to the task of choosing informative patches from unlabeled datasets. Only labeling these patches enables a reduction in manual work of up to 85 percent.

Introduction

Supervised classification plays an important role for extracting semantic information from remote sensing imagery. From statistical considerations, it can be expected that the estimation of any complex model with high accuracy will require large amounts of training data. While unlabeled data are abundant and are already used successfully in unsupervised and semi-supervised learning methods, they cannot completely replace the dependence on labeled data. On the other hand, the acquisition of high quality, densely sampled and representative labeled samples is expensive and a time consuming task. Transfer Learning (TL) is a paradigm that strives to vastly reduce the amount of required training data by utilizing knowledge from related learning tasks (Thrun and Pratt, 1998; Pan and Yang, 2010). In particular, the aim of TL is to adapt a classifier trained on data from a *source domain* to a *target domain*. The only assumption to be made is that these domains are different but related. We are interested in one specific setting of TL called domain adaptation (DA). DA methods assume the source and target domains to differ only by the marginal distributions of the features and the posterior class distributions (Bruzzone and Marconcini, 2009). The performance of DA depends on how the source is related to the target (Eaton *et al.*, 2008). From that point of view, DA can be divided into two steps: find the most similar sources and transfer knowledge from these sources to the target. In this context, the major challenge in source selection is how to measure the similarity of domains.

In this paper, we will address the problems of searching for similar sources, also known as *source selection*, and of integrating the results into DA. As unlabeled data are abundant, our proposed method is only based on similarity measurements between the marginal distributions of the features in the source and target domains. We apply our source selection method to two different data acquisition settings: domain selection and domain ranking. In *domain selection*, given a target domain and a list of candidate source domains, we assign weights to these sources based on the *Maximum Mean Discrepancy* (MMD) metric to the target. For these candidate source domains, we assume that some labeled training data is available from earlier surveys. We then apply *multi-source selection* by transferring knowledge from multiple weighted source domains simultaneously. Additionally, we extend the approach for DA presented in (Paul *et al.*, 2016) so that it can benefit from multi-source selection. For the *domain ranking* setting, we have to process many initially unlabeled target domains while no training data is available. Using our multi-source selection algorithm, our goal is to rank these domains in terms of their informativeness. This information helps us to select the most important domains for manual labeling, which leads to a reduced effort for the generation of training data while keeping classification error at an acceptable level. Finally, we propose an improvement of the MMD metric for the application in source selection with many candidate sources. This Asymmetric Maximum Mean Discrepancy is able to significantly reduce the memory footprint for each source while featuring a linear runtime complexity by exploiting the asymmetric relationship between target and source domains. We evaluate our methods on the Vaihingen and Potsdam datasets from the ISPRS 2D semantic labeing challenge (Wegner *et al.*, 2016) and on a third, even more challenging, dataset based on aerial imagery of three German cities.

Related Work

In our work, we use notation according to Pan and Yang (2010). A domain $\mathcal{D}=\{\mathcal{X}, P(X)\}$ consists of a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$ with $X\in\mathcal{X}$. A task for a given domain is defined as $\mathcal{T}=\{\mathcal{C}, h(\cdot)\}$, consisting of a label space $\mathcal{C}$ and a predictive function $h(\cdot)$. The predictive function can be learned from the training data $\{\mathbf{x}_r, C_r\}$, where $\mathbf{x}_r\in X$ and $C_r\in\mathcal{C}$. We consider a target T, for which we want to learn a predictive function $h(\mathbf{x})$, and a source S, from which some knowledge can be transferred. Both T and S are fully described by their domains and their tasks. In our work, we consider at least one source domain $\mathcal{D}_S$ and only one target domain $\mathcal{D}_T$ for the *domain selection* setting, and more than one target domain for the *domain ranking* setting. There are different settings of TL. Our focus is on DA, which is a special sub-category of the

Karsten Vogt and Jörn Ostermann are with the Institute für Informationsverarbeitung, Leibniz Universität Hannover (vogt@tnt.uni-hannover.de).

Andreas Paul, Franz Rottensteiner, and Christian Heipke are with the Institute of Photogrammetry and Geoinformation, Leibniz Universität, Hannover.

Photogrammetric Engineering & Remote Sensing
Vol. 84, No. 5, May 2018, pp. 249–261.
0099-1112/18/249–261

doi: 10.14358/PERS.84.5.249

transductive TL setting (Pan and Yang, 2010). There are slightly different definitions of the DA problem; we follow the definition of Bruzzone and Marconcini (2009) according to which different domains only differ by the marginal distributions of the features and the posterior class distributions, i.e., we assume $P(X_S) \neq P(X_T)$ and $P(C_S \mid X_S) \neq P(C_T \mid X_T)$. From that point of view, DA corresponds to a problem where the source and target domain data are different, e.g., due to different lighting conditions or seasonal effects. However, the domains must be related, i.e., these differences must not be so large that transfer becomes impossible. In this scenario, finding a solution to the DA problem would allow to transfer a classifier trained on one set of images where training data are available ($\mathcal{D}_S$) to other images ($\mathcal{D}_T$) without having to provide additional training data in $\mathcal{D}_T$. This is different from the problem that the training set is non-representative, e.g., due to class imbalance. Such algorithms are known as *sample selection bias* or *covariate shift* correcting methods, as in (Zadrozny, 2004; Sugiyama *et al.*, 2007). Zhang *et al.* (2010) adapted the classifier to the distribution of the target data by weighing training samples with a probability ratio of data from the source and target domains. However, this approach only deals with binary problems and applications other than image classification.

Pan and Yang (2010) subdivide DA into two groups according to what is actually transferred: *feature representation transfer* and *instance transfer*. Methods of the first group using *feature representation transfer* assume that the differences between domains can be mitigated by projecting both domains into a shared feature space in which the differences between the marginal feature distributions are minimized, e.g., by using feature selection (Gopalan *et al.*, 2011) or feature extraction (Matasci *et al.*, 2015). Some of the methods in this group employ a graph matching procedure to find correspondences between domains (Tuia *et al.*, 2013; Banerjee *et al.*, 2015). These methods need to contain the correct matching sequence among the possible matches or labeled samples across domains to perform well. Cheng and Pan (2014) propose a semisupervised method for DA that uses linear transformations for feature representation transfer. However, this method also requires training data from the target domain. Methods that assume that differences can be found in the marginal distributions mostly fall into the second group of DA algorithms, based on *instance transfer*. These methods try to directly refuse training samples from the source domain, successively replacing them by samples from the target domain that receive their class labels (*semi-labels*) from the current state of the classifier.

Methods for instance transfer have been used in the classification of remotely sensed data, e.g., in (Acharya *et al.*, 2011). Acharya *et al.* (2011) train the classifier on the basis of the source domain and combine the result with those of several clustering algorithms to obtain improved posterior probabilities for the target domain data. The approach is based on the assumption that the data points of a cluster in feature space probably belong to the same class. Bruzzone and Marconcini (2009) present a method for DA based on instance transfer for Support Vector Machines (SVM). In Paul *et al.* (2016), this idea was adapted to logistic regression, which has a lower computational complexity in training for multiclass problems. Durbha *et al.*, (2011) show that methods of TL for classification of remotely sensed images can produce better results than a modification of the SVM. A DA method using logistic regression in a semi-supervised setting combined with clustering of unlabeled data has been presented in (Amini and Gallinari, 2002). Training is based on expectation maximization (EM), and the semi-labels of the unlabeled data are determined according to the cluster membership of EM. In contrast to our DA technique, that method assumes the labeled and the unlabeled data to follow the same distribution.

The detection of negative transfer is of vital importance for TL. In (Bruzzone and Marconcini, 2010) a circular validation scheme was proposed to detect negative transfer after adapting the classifier. An alternative approach, *source selection*, would try to detect a relevant source prior to applying TL, which, of course, requires the availability of multiple source domains. Most work in this area uses a distance measure between the marginal distributions to measure the similarity between domains. Such distribution distances are well known in statistics, where the problem is mostly solved for 1D feature spaces. Most research has therefore focused on extending these metrics to multivariate data by using non-parametric models. Examples for such measures are the *Kullback-Leibler Divergence* (Sugiyama *et al.*, 2007), the *Total-Variation Distance* (Sriperumbudur *et al.*, 2012) and its approximations, the *Maximum-Mean-Discrepancy* (Gretton *et al.*, 2012; Chattopadhyay *et al.*, 2012; Matasci *et al.*, 2015) and *A-Distance* (Ben-David *et al.*, 2007). These approaches are kernel-based and usually scale well to high-dimensional data, but they may be computationally expensive. Therefore, another focus of research has been on reducing computational requirements and an improved regularization by careful kernel tuning (Zaremba *et al.*, 2013; Sriperumbudur *et al.*, 2009).

Chattopadhyay *et al.* (2012) proposed a multi-source DA algorithm for the detection of muscle fatigue from surface electromyography (SEMG) data. The data show a high variability between individual subjects, therefore not all subject data should be considered when learning an individualized fatigue detector for a new subject. A synthesized source is generated as a weighted combination of all candidate sources using a MMD-based domain distance. The method has cubic complexity in the number of candidate sources, which may make it slow for cases with many available sources.

Besides TL, *active learning* has also been an active research topic with the aim to reduce manual labeling costs (Settles, 2010). *Active learning* methods select the most informative samples from an initially unlabeled training set which are presented to a human operator for labeling. Further samples may then be selected while taking the user feedback and the peculiarities of the classifier into account. Some ideas were proposed to utilize active learning for DA (Tuia *et al.*, 2011). While our *domain ranking* setting bears some similarity to *active learning*, our approach works at a coarser level and does not incorporate a user feedback loop, resulting in a much simpler user workflow and faster computation times.

In this paper, we present an unsupervised and a supervised method for source selection based on different distance metrics for domains. The work is inspired by (Chattopadhyay *et al.*, 2012), but we use an approximate optimization with linear run-time complexity and propose a method for tuning the kernel hyperparameter automatically. The methods deliver a synthetic source as a weighted combination of similar sources, designed to reduce a distance between the distributions of the synthetic source and the target domains. We also propose variations of our distance metrics that are able to exploit the asymmetrical relationship between target and source domains in TL. Furthermore, we extend the algorithm in (Paul *et al.*, 2016) so that it can deal with multiple sources. Finally, we apply our proposed methods to rank a set of target domains in order of their informativeness. Only the most informative domains need to be labeled manually in order to generate high quality semantic segmentation for all targets.

Domain Adaptation

We start this section with a short description of the work from (Paul *et al.*, 2016) before presenting our improvements in the next section.

DA Approach

We use multiclass *logistic regression* (LR) as our base classifier. LR directly models the posterior probability $P(C \mid \mathbf{x})$ of the class labels C given the data $\mathbf{x}$. We transform features into a higher-dimensional space $\Phi(\mathbf{x})$ in order to be able to achieve non-linear decision boundaries. In the multiclass case, the model of the posterior is based on the Softmax function (Bishop, 2006):

$$P\left(C = C^k \mid \mathbf{x}\right) = \frac{exp\left(\mathbf{w}_k^T \cdot \Phi(\mathbf{x})\right)}{\sum_j exp\left(\mathbf{w}_j^T \cdot \Phi(\mathbf{x})\right)} \quad (1)$$

where $\mathbf{w}_k$ is a parameter vector for a particular class label C^k to be determined in the training process for the class $k \in K$. For that purpose, a *training data* set, denoted as $\overline{TD}$ is assumed to be available. Initially, it contains only training samples from the source domain, each consisting of a feature vector $\mathbf{x}_n$, its class label C_n and a weight g_n. In the initial training, we use $g_n = 1$ for each sample $n \in \{1, \ldots, N\}$, but in the DA process, the samples will receive individual weights indicating the algorithm's confidence in the labels. In training, the optimal values of the parameter vector $\mathbf{w}$ (collecting the parameter vectors $\mathbf{w}_k$ for all classes k) given $\overline{TD}$ are determined by optimizing the posterior (Vishwanathan *et al.*, 2006):

$$p\left(\mathbf{w} \mid \overline{TD}\right) \propto p(\mathbf{w}) \cdot \prod_{n,k} p\left(C_n = C^k \mid \mathbf{x}_n, \mathbf{w}\right)^{g_n \cdot q_{nk}} \quad (2)$$

where q_{nk} is 1 if $C_n = C^k$ and 0 otherwise, $p(C = C^k \mid \mathbf{x}_n, \mathbf{w})$ is defined in Eq. (1) and $p(\mathbf{w})$ is a Gaussian prior with mean $\bar{\mathbf{w}}$ and standard deviation σ. Compared to standard multiclass LR, the only difference is the use of the weights g_n (Paul *et al.*, 2016). We use the Newton-Raphson method for finding the optimal parameters $\mathbf{w}$ by minimizing $-log(p(\mathbf{w} \mid \overline{TD}))$ (Bishop, 2006).

Our aim is to transfer the classifier trained on labeled source domain data to the target domain in an iterative procedure. Our initial classifier is trained on the training set $\overline{TD}^0$ containing only source data. In each further iteration i of DA a predefined number ρ_E of source samples is removed from and a number ρ_A of semi-labeled target samples is included into the current training data set $\overline{TD}^i$. Thus, in iteration i, the current training data set $\overline{TD}^i$ consists of a mixture of N_S^i source samples and N_T^i target samples:

$$\overline{TD}^i = \left\{\left(\mathbf{x}_{S,r}; C_{S,r}; g_{S,r}\right)\right\}_{r=1}^{N_S^i} \bigcup \left\{\left(\mathbf{x}_{T,l}; \tilde{C}_{T,l}; g_{T,l}\right)\right\}_{l=1}^{N_T^i}.$$

The symbol $\tilde{C}_{T,l}$ denotes the *semi-labels* of the target samples, which are determined by applying a criterion based on the class labels of the k nearest neighbors (*knn*) of a sample in feature space. If the most frequent class label among the *knn* of an unlabeled sample is consistent with the predicted label according to a current state of the LR classifier, it is considered a candidate for inclusion into $\overline{TD}^i$. The ρ_A candidate samples having the shortest average distance to their k nearest neighbors will be added to $\overline{TD}^i$. We first remove source samples that are most distant from the decision boundary starting with the samples showing inconsistent class labels and continuing with samples with consistent labels. As i is increased, N_S^i becomes smaller and N_T^i increases, until finally, only target samples with semi-labels are used for training.

At each iteration i, we have to define sample weights $g_{\overline{TD}}^i \in [0,1]$ for all training samples in $\overline{TD}^i$, where $\left\{g_{\overline{TD}}^i\right\} = \left\{\left\{g_{S,r}^i\right\}_{r=1}^{N_S^i} \bigcup \left\{g_{T,l}^i\right\}_{l=1}^{N_T^i}\right\}$. For simplicity, we refer to the weight of a sample as $g_{(\overline{TD},n)}^i$, $n \in \{1, .., N^i\}$ with $N^i = |\overline{TD}^i|$ the number of elements in that training set if it does not matter whether the sample is originally from the source or from the target domain. The weight indicates the algorithm's trust in the correctness of the label of a training sample. The weight function used for determining $g_{(\overline{TD},n)}^i$ depends on the distance to the decision boundary: the higher that distance, the higher is the weight; a parameter h models the rate of increase of the weight with the distance (Paul *et al.*, 2016). Having defined the current training data set $\overline{TD}^i$ and the weights, we retrain the LR classifier. This leads to an updated parameter vector w and a change in the decision boundary. This new state of the classifier is the basis for the definition of the training data set in the next iteration. Thus, we gradually adapt the classifier to the distribution of the target data.

Multi-Source Logistic Regression DA

In this section, the method previously described is adapted for using data from multiple source domains for training. To formally state our problem, we define our current training data set as follows:

$$\overline{TD}^i = \bigcup_{s=1}^{|\mathbb{S}|} \left\{\left(\mathbf{x}_{S^s,r}; C_{S^s,r}; g_{S^s,r}\right)\right\}_{r=1}^{N_{S^s}^i} \bigcup \bigcup_{t=1}^{|\mathbb{T}|} \left\{\left(\mathbf{x}_{T^t,l}; C_{T^t,l}; g_{T^t,l}\right)\right\}_{l=1}^{N_{T^t}^i}, \quad (3)$$

where $\mathbb{S}$ or $\mathbb{T}$ describe a set of source or target data sets, respectively, and $|\mathbb{T}| = 1$.

Again, we refer to a particular sample in $\overline{TD}^i$ by its index n in $\overline{TD}^i$ if we are not interested in the domain it comes from. We use the defined training data set $\overline{TD}^i$ in our multi-source DA approach, but we use different definitions of the sample weights. One modification of sample weights should decrease the weight of uncertain samples; the other one is required to deal with prior weights assigned to the individual source domains (See the next Section).

Sample Weights

The individual weights for the training samples should indicate the algorithm's trust in the correctness of the semi-labels, but the definition of weights in (Paul *et al.*, 2016) only depended on the distance of a sample from the decision boundary. It may happen that a semi-label changes in the iterative DA process, which would imply that the semi-label is uncertain; semi-labels not having changed for many iterations should be trusted more than others. Here, we introduce an adapted definition of the sample weights as shown in (Chang *et al.*, 2002; Bruzzone and Marconcini, 2009; Matasci *et al.*, 2012) to model the trust in a sample in $\overline{TD}$ as a function of the number of iterations j for which its semi-label has remained unchanged (Figure 1):

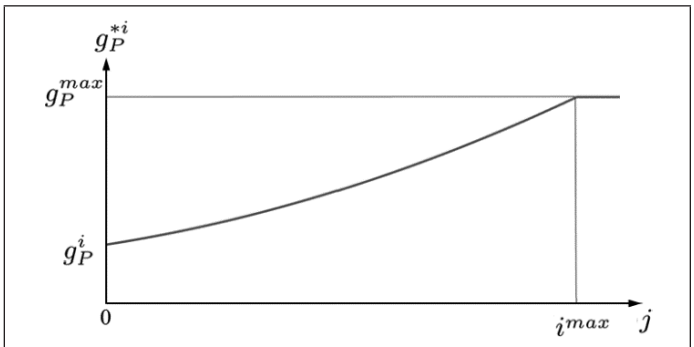


Figure 1. Sample weight function according to Eq. (4), assuming constant g_n^i during the adaptation.

$$g_n^{*i} = min\left(g_n^i + \frac{\left(g^{max} - g_n^i\right)\cdot j^2}{\left(i^{max}-1\right)^2}, g^{max} \right). \quad (4)$$

In Equation 4, g_n^i is the weight of sample n in the current adaptation step i according to the original distance-based weight function (Paul *et al.*, 2016), g_n^{*i} is the new weight of that sample, i^{max} defines the number of iterations for which the weight of a samples is allowed to increase quadratically with j, and g^{max} is the maximum possible sample weight. If only one source domain were considered, the weight for each training sample n in $\overline{TD}^i$ would be g_n^{*i}, i.e., the algorithm outlined in the previous Section would be applied using the new definition of weights.

Domain Weights

In the context of multi-source selection, we introduce an individual domain weight π_{S^s} for every source domain s used in the DA process. The domain weights allow us to obtain a synthesized source $\overline{S}$ (See the next Section) from multiple sources that is more similar to the target domain than any of the original ones. The domain weights remain constant during the adaptation procedure. For a sample n in the current training set $\overline{TD}^i$ taken from source domain s, the weight used in the DA process is $g_{\overline{TD},n}^i = g_n^{*i} \cdot \pi_{S^s}$, where g_n^{*i} is defined in Equation 4, whereas a sample n with a semi-label taken from the target domain has only the weight g_n^{*i}. Thus, the weights of the source-domain samples are affected by the similarity of the corresponding domain to the target domain, placing a higher trust into samples that come from more similar source domains.

Multi-Source Selection

The goal of source selection is to improve the prospects of DA by choosing a source $\overline{S}$ that is, in some sense, most similar to the target domain. Naturally, one should prefer sources that produce similar decision boundaries as the target task. Therefore, the selection criterion should be based on $\varepsilon(h_S,\overline{TD}_T)$, i.e., the relative classification error ($\in[0,1]$) on the target data, given the predictive function h_S of the source task:

$$\overline{S} = \text{argmin}_{S\in\mathbb{S}}\varepsilon\left(h_S,\overline{TD}_T\right) \quad (5)$$

The main difficulty lies in the fact that estimating the classification error requires the class labels of the target domain to be known. Here, we introduce a theoretical framework and outline an algorithm that allows us to quickly find approximate solutions while requiring much less information. We first design two complementary domain distance functions, which we call d_{SDA} and d_{UDA}. The function d_{SDA} measures a supervised domain distance in the sense that only class labels in the source domain need to be known, whereas d_{UDA} does not require any class labels at all. We refer to d_{DA} in places where either of these functions could be used. Equation 5 can then be approximated by $\overline{S} = \text{argmin}_{S\in\mathbb{S}} d_{\text{DA}}(.)$. Our main contribution is the extension of these domain distances to the transfer from multiple sources while having a linear run-time complexity. In addition, we also developed variants of these domain distances that are able to capture the often asymmetric relationship between the target and source domains in TL. Finally, we also show how all critical hyperparameters can be tuned automatically in an efficient manner.

Similarity of Domains

We derive our approximation of Equation 5 in several steps. Using the results of Ben-David *et al.* (2007), an upper bound for the classification error can be given as:

$$\varepsilon\left(h_S,\overline{TD}_T\right) \le \varepsilon\left(h_S,\overline{TD}_S\right) + d_{\mathcal{A}}\left(\overline{TD}_T,\overline{TD}_S\right) + \gamma \quad (6)$$

The first term corresponds to the classification error on the source task. The term $d_{\mathcal{A}}(\overline{TD}_T,\overline{TD}_S)$, called $\mathcal{A}$-distance, describes a distance between the marginal feature distributions of the source and target domains. The third term, γ, encapsulates to which degree the DA assumption holds. The exact value can only be computed if class labels in the target task are available, but for related datasets, this term should only take small positive values. Assuming that γ is unknown yet constant over the dataset, the upper bound gives us a definition for d_{SDA} according to $d_{\text{SDA}} = \varepsilon(h_S,\overline{TD}_S) + d_{\text{A}}(\overline{TD}_T,\overline{TD}_S)$. In the following, we define $d_{\mathcal{A}}$ and derive a more computationally friendly way to estimate this distribution distance. In (Ben-David *et al.*, 2007), the A-distance is defined as:

$$d_{\mathcal{A}}(\overline{TD}_T,\overline{TD}_S) = 2(1-2\varepsilon(h_{T\perp S},\overline{TD}_{T\perp S})) \quad (7)$$

The term $\varepsilon(h_{T\perp S},\overline{TD}_{T\perp S})$ describes the classification error for a classifier discriminating between feature vectors from the source and target domains. In the referenced paper, only signed linear classifiers such as SVMs or logistic regression models were considered. Evaluation of the $\mathcal{A}$-distance involves the training of such a classifier for each candidate source, which has a high computational complexity. Furthermore, linear separability of the source and target domains is explicitly assumed. It is therefore desirable to find an approximation to the $\mathcal{A}$-distance that displays more favorable properties. Gretton *et al.* (2012) independently proposed the Maximum Mean Discrepancy (MMD) as a general distance function between probability distributions:

$$\begin{aligned} d_{MMD}^2\left(\overline{TD}_T,\overline{TD}_S\right) &= E\left[\left(\phi\left(\boldsymbol{x}_T\right)-\phi\left(\boldsymbol{x}_S\right)\right)^2\right] \\ &= E\left[k(\boldsymbol{x}_T,\boldsymbol{x'}_T)\right] - 2E\left[k\left(\boldsymbol{x}_T,\boldsymbol{x}_S\right)\right] + E\left[k\left(\boldsymbol{x}_S,\boldsymbol{x'}_S\right)\right] \end{aligned} \quad (8)$$

where $\boldsymbol{x}$ and $\boldsymbol{x}'$ are statistically independent samples from the same distribution. The MMD computes the distance between the means of the probability distributions in a *Reproducing Hilbert Kernel Space* (RKHS). The RKHS is uniquely defined by either a feature space mapping $\phi(\boldsymbol{x})$ or its kernel function $k(\boldsymbol{x},\boldsymbol{y})$. It was shown by Sriperumbudur *et al.* (2012) that the relation

$$d_{\mathcal{A}}(\overline{TD}_T,\overline{TD}_S) \approx 2d_{MMD}(\overline{TD}_T,\overline{TD}_S) \quad (9)$$

holds for positive bounded kernels such as the Gaussian kernel:

$$k_{RBF}\left(\boldsymbol{x},\boldsymbol{y}\right) = \exp\left(-\frac{\boldsymbol{x}-\boldsymbol{y}^2}{2\sigma^2}\right) \quad (10)$$

Evaluation of the MMD can be done by replacing the expectations in Equation 8 with their empirical estimates. A naive estimator would have a run-time complexity of $\mathcal{O}(N_T\cdot N_S)$, where N_T and N_S are the numbers of features available in the target and source domains, respectively, which becomes untenable for large training sets. A much faster linear-time estimator d_{LMMD} was proposed by Gretton *et al.* (2012). Assuming $M = N_T = N_S$, it can be stated as:

$$\begin{aligned} &d_{LMMD}^2\left(\overline{TD}_T,\overline{TD}_S\right) \\ &= \frac{2}{M}\left[\sum_{r=1}^{M/2} k(\boldsymbol{x}_{T,2r},\boldsymbol{x}_{T,2r-1}) - \sum_{r=1}^{M} k(\boldsymbol{x}_{T,r},\boldsymbol{x}_{S,r}) + \sum_{r=1}^{M/2} k(\boldsymbol{x}_{S,2r},\boldsymbol{x}_{S,2r-1})\right] \end{aligned} \quad (11)$$

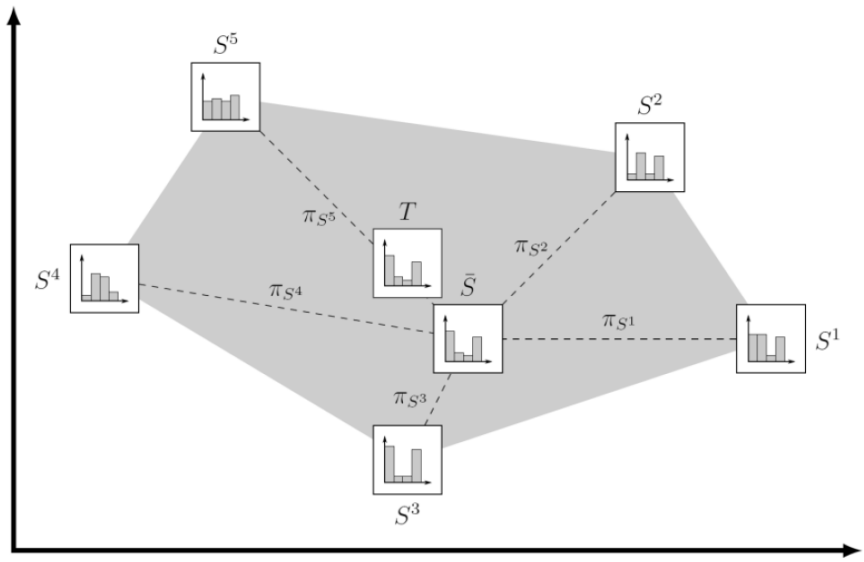


Figure 2. A synthesized source $\bar{\boldsymbol{S}}$ is formed as the convex combination of candidate sources $\boldsymbol{S}^s$.

Finally, replacing $d_{\mathcal{A}}$ by d_{LMMD} using Equation 9 leads to the definition of our supervised domain distance:

$$d_{SDA}(\overline{TD}_T, \overline{TD}_S) = \varepsilon\left(h_S, \overline{TD}_S\right) + 2d_{LMMD}(\overline{TD}_T, \overline{TD}_S) \quad (12)$$

Assuming the classification error to be approximately constant over all candidate sources, we obtain the unsupervised distance:

$$d_{UDA}\left(\overline{TD}_T, \overline{TD}_S\right) = 2d_{LMMD}\left(\overline{TD}_T, \overline{TD}_S\right) \quad (13)$$

Asymmetric Domain Distance

The described domain distances based on the MMD, see Equations 12 and 13, are theoretically motivated and assume a symmetric relationship between the source and target domains, e.g., if a classifier learned from $\overline{TD}_S$ performs well on $\overline{TD}_T$, then the reverse must also be true. In reality, this assumption may not always hold. For instance if $\overline{TD}_T \subset \overline{TD}_S$ we should expect that a classifier learned on S will perform well on T as all classes are well represented by the training data. Yet, the distributions are measurably different, which can be observed by a high MMD measure. We therefore propose a modification of the MMD which is aimed at directly measuring whether all regions in T are represented in S, while being invariant to those regions in S that are conversely not represented in T. First, let us re-examine the MMD from Equation 8:

$$\begin{aligned} d^2_{MMD}\left(\overline{TD}_T, \overline{TD}_S\right) &= E\left[k(\boldsymbol{x}_T, \boldsymbol{x'}_T)\right] - 2E\left[k\left(\boldsymbol{x}_T, \boldsymbol{x}_S\right)\right] + E\left[k\left(\boldsymbol{x}_S, \boldsymbol{x'}_S\right)\right] \\ &= \underbrace{\left(E\left[k(\boldsymbol{x}_T, \boldsymbol{x'}_T)\right] - E\left[k(\boldsymbol{x}_T, \boldsymbol{x}_S)\right]\right)}_{Target} \\ &+ \underbrace{\left(E\left[k(\boldsymbol{x}_S, \boldsymbol{x'}_S)\right] - E\left[k(\boldsymbol{x}_S, \boldsymbol{x}_T)\right]\right)}_{Source} \end{aligned} \quad (14)$$

It should be noted that this is valid for the Gaussian kernel since $k_{\text{RBF}}(x,y) = k_{\text{RBF}}(y,x)$. The term $E[k(\boldsymbol{x}_S, \boldsymbol{x'}_S]$ describes the compactness of the source domain, which is mostly irrelevant for measuring the relatedness of a source. We therefore drop the entire second part of this equation. The remaining terms describe the average intra domain similarity ($T \Leftrightarrow T$) and inter domain similarity ($T \Leftrightarrow S$), respectively. We argue that for each sample in the target domain to be well represented, a related source should contain at least one sample that is not significantly more dissimilar than its next most similar target sample. Therefore, we propose to replace the simple average in the MMD with a maximum operator over similarity scores:

$$\begin{aligned} d^2_{AMMD}\left(\overline{TD}_T, \overline{TD}_S\right) &= \frac{1}{N_T}\sum_{i=1}^{N_T} \max_{j \in [1..N_T]} k\left(\boldsymbol{x}_{T,i}, \boldsymbol{x'}_{T,j}\right) \\ &- \frac{1}{N_T}\sum_{i=1}^{N_T} \max_{j \in [1..N_S]} k\left(\boldsymbol{x}_{T,i}, \boldsymbol{x}_{S,j}\right) \end{aligned} \quad (15)$$

A disadvantage of this formulation is that in order to find the most similar sample we always have to look at the entire training set. Therefore, Equation 15 has a quadratic run-time complexity. Yet, suppose we are content with finding only the q% most similar samples and we further also allow a failure probability p for locating such a sample. Then, it can be shown that it is sufficient to only look at a random subset of size $N_{\max} \geq \log_{(100\%-q)}(p)$, irrespective of the underlying data distribution or sample size (See Appendix A for the proof). Consequently, Equation 15 can be approximated without significant loss of accuracy using a procedure having linear run-time complexity. This result also has the benefit that a source selection system based on our AMMD never has to store the full source training sets to perform queries. In fact, for each source only $N_{\max}$ samples have to be held in memory, where $N_{\max}$ is typically less than 100. Using these results, one can use the metric d_{AMMD} to obtain modified (asymmetric) versions of the domain distances d_{SDA} and d_{UDA}:

$$d_{A-SDA}(\overline{TD}_T, \overline{TD}_S) = \varepsilon\left(h_S, \overline{TD}_S\right) + 2d_{AMMD}(\overline{TD}_T, \overline{TD}_S) \quad (16)$$

$$d_{A-UDA}(\overline{TD}_T, \overline{TD}_S) = 2d_{AMMD}(\overline{TD}_T, \overline{TD}_S) \quad (17)$$

Convex Combination of Domains

In general, we have to expect that none of the candidate source domains $S \in \mathbb{S}$ is a perfect match for the target domain. Nonetheless, the target marginal distribution $p_T(\boldsymbol{x})$ might be much closer to the subspace spanned by the convex combination of the source marginal distributions (Figure 2). Any point

in this subspace represents a valid marginal distribution and can be parametrized as:

$$p_{S_\pi}(\boldsymbol{x}) = \sum_{s=1}^{|\mathbb{S}|} \pi_{S^s} p_{S^s}(\boldsymbol{x}) \tag{18}$$

given a source weight vector $\pi = \left[\pi_{S^1}, \ldots, \pi_{S^{|\mathbb{S}|}}\right]^T$ satisfying the constraints $\pi_{S^s} \geq 0, \sum_{s=1}^{|\mathbb{S}|} \pi_{S^s} = 1$. By definition (Equation 18), the distribution $p_{S_\pi}(\boldsymbol{x})$ is a mixture of the source marginal distributions. The weighted training set $\overline{TD}_{S_\pi} = \bigcup_{s=1}^{|\mathbb{S}|} \left\{\boldsymbol{x}_{S^s}; C_{S^s}; \pi_{S^s}\right\}_{r=1}^{N_{S^s}}$ is therefore a representative sample of this distribution. The weights can be intuitively understood to mean that each sample from source $S^s \in \mathbb{S}$ is counted as π_{S^s} such samples. As an important intermediate result, we propose extensions of the linear-time MMD estimator (Equation 11) and our asymmetric MMD (Equation 15) to a weighted union of source training sets:

$$\begin{aligned} & d^2_{LMMD}\left(\overline{TD}_T, \overline{TD}_{S_\pi}\right) \\ &= \frac{2}{M}\left[\sum_{r=1}^{M/2} k\left(\boldsymbol{x}_{T,2r}, \boldsymbol{x}_{T,2r-1}\right) - \sum_{u=1}^{|\mathbb{S}|} \pi_{S^u} \sum_{r=1}^{M} k\left(\boldsymbol{x}_{T,r}, \boldsymbol{x}_{S^u,r}\right)\right. \\ &\quad + \sum_{u=1}^{|\mathbb{S}|} \sum_{v=u+1}^{|\mathbb{S}|} \pi_{S^u} \pi_{S^v} \sum_{r=1}^{M} k\left(\boldsymbol{x}_{S^u,r}, \boldsymbol{x}_{S^v,r}\right) \\ &\quad \left. + \sum_{u=1}^{|\mathbb{S}|} \pi^2_{S^u} \sum_{r=1}^{M/2} k\left(\boldsymbol{x}_{S^u,2r}, \boldsymbol{x}_{S^u,2r-1}\right)\right] \end{aligned} \tag{19}$$

$$\begin{aligned} & d^2_{AMMD}\left(\overline{TD}_T, \overline{TD}_{S_\pi}\right) = \frac{1}{N_T} \sum_{i=1}^{N_T} \max_{j \in [1..N_T]} k\left(\boldsymbol{x}_{T,i}, \boldsymbol{x'}_{T,j}\right) \\ & - \frac{1}{N_T} \sum_{u=1}^{|\mathbb{S}|} \pi_{S^u} \sum_{i=1}^{N_T} \max_{j \in \left[1..N_{S^u}\right]} k\left(\boldsymbol{x}_{T,i}, \boldsymbol{x}_{S^u,j}\right) \end{aligned} \tag{20}$$

In the next section, we present a fast and greedy optimization scheme that minimizes d_{DA} w.r.t. π.

Fast Synthesis of Source Domains by Boosting

Convex representation problems, like the one in Equation 18, are related to dictionary learning. The Iterative Nearest Neighbor (INN) algorithm (Timofte and Van Gool, 2012) is a recent method that approximately solves such problems in a greedy fashion. The solution at iteration L is given as:

$$p^L_S(\boldsymbol{x}) = \sum_{l=1}^{L} w^l p_{S_l}(\boldsymbol{x}), \tag{21}$$

where the iteration weights are computed as:

$$w^l = \frac{\lambda}{(1+\lambda)^l} \tag{22}$$

for a fixed parameter λ. In order to find the next solution $p^{L+1}_S(\boldsymbol{x})$, we select a source which minimizes the representation error to the target domain according to our domain distance:

$$S_{L+1} = \underset{S \in \mathbb{S}}{\operatorname{argmin}}\, d_{DA}\left(\overline{TD}_T, \left\{\boldsymbol{x}_{S,r}; C_{S,r}; w^{L+1}\right\}_{r=1}^{N_S} + \bigcup_{l=1}^{L} \left\{\boldsymbol{x}_{S_l,r}; C_{S_l,r}; w^l\right\}_{r=1}^{N_{S_l}}\right) \tag{23}$$

The same source may be chosen multiple times at different iterations. The source weights can be derived from the iteration weights as follows:

$$\pi_{S^s} = \sum_{l=1}^{L} w^l \cdot 1_{\left\{S_l = S^s\right\}}. \tag{24}$$

Originally, the INN algorithm was designed to work on vectors in Euclidean spaces. When interpreted in the space of probability distributions, the procedure has strong parallels to a non-adaptive variant of the boosting paradigm, whose most well-known implementation is AdaBoost (Schapire and Singer, 1999). Similar to boosting, the synthesized source S_π is a weighted combination of weaker approximations. In addition, the update step in Equation 23 has the effect to steer the optimization successively to prioritize parts of the distribution that are not yet well represented while also attenuating overrepresented parts.

The sum $\sum_{l=1}^{\infty} w^l$ approaches 1 while the iteration weights w^l become smaller and smaller. We can therefore stop the algorithm after L iterations such that $\sum_{l=1}^{L} w^l > \beta$ while avoiding large approximation errors. From Equation 22 follows:

$$L = -\frac{\log(1-\beta)}{\log(1+\lambda)} \tag{25}$$

For typical parameter values $\beta = 0.9$, $\lambda = 0.5$ only, and $L = 6$ required iterations. The run-time complexity of the entire multi-source selection algorithm using $d_{\{\mathrm{UDA,A\text{-}UDA}\}}$ can be given as $\mathcal{O}(L^3 \cdot |\mathbb{S}| \cdot M)$. The same result for our supervised variants $d_{\{\mathrm{SDA,A\text{-}SDA}\}}$ reads as $\mathcal{O}(L^3 \cdot |\mathbb{S}| \cdot M \cdot f(|\mathbb{S}| \cdot M))$ and additionally depends on the term $f(|\mathbb{S}| \cdot M)$, which describes the complexity of the classification algorithm used to estimate the first term in Equation 12.

Algorithm 1 Kernel Bandwidth Estimation

$\phi \leftarrow 1.61803398875$
$(L,R) \leftarrow (0, \pi/2)$
$(A,B) \leftarrow (R - (R-L)/\phi,\ L + (R-L)/\phi)$
for i = 1..MaxIter **do**
  $f_A \leftarrow d^2_{MMD}(\overline{TD}_T, \overline{TD}_S)$ with $\sigma = \tan(A)$
  $f_B \leftarrow d^2_{MMD}(\overline{TD}_T, \overline{TD}_S)$ with $\sigma = \tan(B)$
  if $f_A < 0$ **then**
    $R \leftarrow A$
  else if $f_B \leq f_A$ **then**
    $R \leftarrow B$
  else
    $L \leftarrow A$
  end if
  $(A,B) \leftarrow (R - (R-L)/\phi,\ L + (R-L)/\phi)$
end for
return $\sigma_{\max} = \tan((L+R)/2)$

Kernel Bandwidth Estimation

The Gaussian kernel has a single hyperparameter σ, its bandwidth. It was shown by Sriperumbudur *et al.* (2009) that the discriminative power of the MMD is maximized by maximizing d_{MMD} with respect to σ:

$$d^2_{\overline{MMD}}\left(\overline{TD}_T, \overline{TD}_S\right) = \max_{\sigma\in(0,\infty)} d^2_{MMD}\left(\overline{TD}_T, \overline{TD}_S\right) \quad (26)$$

Using the results by Shestopaloff (2010), we can show that this optimization problem has exactly one maximum at σ_{max} and at most one minimum at σ_{min}. Furthermore, if σ_{min} exists then $\sigma_{max} < \sigma_{min}$ holds. Finally, d_{MMD} will tend towards zero for both $\sigma\to 0$ and $\sigma\to\infty$. We can therefore conclude that $d_{MMD}(\sigma_{min})<0$ if a minimum exists. Whereas theoretically, the MMD only can take positive values, this case can still occur for very similar domains due to errors in the empirical estimates of the expectations. The general shape of the function $d_{\overline{MMD}}$ is shown in Figure 3.

We solve this optimization problem using a *Golden-Section-Search* (GSS) (Press, 2007) (see Algorithm 1). The GSS searches the maximum of a strictly unimodal function. We modified the GSS to handle the case where a minimum σ_{min} exists. The value range $(0,\infty)$ is mapped to $(0,\pi/2)$ using the *atan* function. In our experiments, the algorithm typically converged in less than 10 iterations. Our empirical evaluation in the Experiments Section shows that the same approach is also valid for our asymmetrical MMD.

Improving Robustness by Bootstrap Aggregation

As all empirical estimators, our MMD estimators have a non-zero estimation variance which may result in a suboptimal solution $\boldsymbol{\pi}$. We propose to reduce this variance by averaging $\boldsymbol{\pi}$ over multiple independent runs of our multi-source selection algorithm. Each run is performed on a bootstrap sample of the training sets $\overline{TD}_T$ and $\overline{TD}_S$. Bootstrap sampling describes a procedure where a new sample is generated using independent draws with replacement from an input sample. The statistical properties of bootstrap sampling are described in detail in (Hesterberg *et al.*, 2003).

Ranking of Source Domains

The *domain ranking* setting might resemble a more relevant workflow for the supervised classification of remote sensing imagery than source selection as previously presented. We assume that we have to process a batch of E images for which initially no training data is available. In order to create some training data we have to label some of these images manually. Obviously, we do not intend to label all of them. In this setting all images can be considered as target domains $T_e\in\mathbb{T}$, while only some of them will also be used as source domains S_s for our source-selection algorithm. Our goal is to find a small subset S_s that will be sufficient to achieve acceptable classification results. A reasonable workflow could be to label source domains sequentially, training a classifier whenever a new source domain is added and applying that classifier to all target domains; a visual inspection of the results could guide the decision when to stop labeling new domains. A *domain ranking* algorithm must therefore be able to compute an ordering of the domains of the batch in which the most informative domains are placed early. For computational reasons we have chosen to restrict our research to greedy algorithms. We use a variant of the *kernel herding* algorithm by Chen *et al.* (2012). *Kernel herding* greedily selects a small representative *super sample* from a larger sample of an unknown distribution. At each step, we select a new sample that is similar to many other samples while also being dissimilar to already selected ones. Due to its formulation as a kernel method, *kernel herding* is flexible enough to be adapted to the *domain ranking* problem. Only a kernel matrix K needs to be defined which encapsulates a pairwise similarity measure between domains. A simple kernel matrix could be directly constructed from the MMD as $k^{MMD}_{i,j} = 1 - d_{MMD}\left(\overline{TD}_{T_i}, \overline{TD}_{T_j}\right)$. While this simple approach typically produces good results, we have determined empirically that it can be far from the optimum if less than five sources are to be selected. Therefore, we propose a more elaborate method to supersede the simple $k^{MMD}_{i,j}$ domain kernel. We first note that the source weights $\boldsymbol{\pi}$ from our multi-source selection algorithm also describe a domain similarity, as more related sources are associated with larger weights. To construct K we first apply multi-source selection to each T_e using any of our unsupervised domain distances ($d_{[\text{UDA,A-UDA}]}$) while using all other domains as candidate sources. We define the e^{th} column vector of K as the source weight vector $\boldsymbol{\pi}_S$ for the e^{th} domain. We also have to consider self-similarity of domains by setting the main diagonal of K to 1. The *kernel herding* algorithm then starts with an empty set S_s of selected source domains. At each iteration the next most informative source domain S^{select} is chosen as:

$$S^{select} = \underset{T_u\in\mathbb{T}\setminus S_s}{\text{argmax}}\left[\frac{1}{|\mathbb{T}|}\sum_{T_e\in\mathbb{T}} k_{u,e} - \frac{1}{|S_s|+1}\sum_{S_v\in S_s} k_{u,v}\right] \quad (27)$$

and added to S_s. The main result of the algorithm is the order in which datasets should be selected for labeling so that they can serve as source domains.

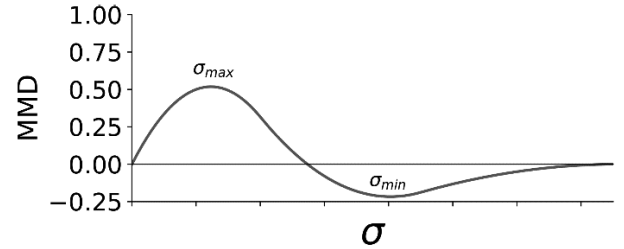


Figure 3. General shape of the optimization problem for hyperparameter estimation. The optimal $\boldsymbol{\sigma}$ can be found at $\boldsymbol{\sigma}_{max}$.

Experiments

Test Data and Test Setup

Our experimental evaluation is based on three datasets (see Figure 4). Two of them are the Vaihingen and Potsdam datasets from the ISPRS 2D semantic labeling contest (Wegner


(a) Vaihingen (b) Potsdam (c) Buxtehude (d) Hannover (e) Nienburg

Figure 4. Details of orthophotos of all datasets used in the experiments.

et al., 2016). The Potsdam dataset was resampled from 5 cm to a ground sampling distance (GSD) of 8 cm to reduce the computational burden. Only patches for which a reference is available were used in our experiments. A third dataset, referred to as 3CITYDS, consists of three regions of German cities of varying size, degree of urbanization, and architecture (Buxtehude, Hannover, Nienburg)[1]. This diversity produces much more pronounced differences between domains. Each region covers an area of 2×2 km^2 but is evenly split up into nine patches. The reference data for the 3CITYDS dataset was generated manually based on the image data. For all datasets, both orthophotos and digital surface models (DSM) generated by image matching are available. The properties of all datasets are given in Table 1. Finally, we also consider a fourth dataset, *Combined*, which is the union of the Vaihingen, Potsdam, and 3CITYDS datasets.

All experiments are based on a pixel-wise classification of the input data into the four object classes *building, tree, low vegetation,* and *impervious surface*. The *impervious surface* class also includes clutter and cars. Furthermore, we used the same feature space for all datasets. Under this constraint, we selected the five most discriminative features using a *Random Forest*-based feature selection method (Breiman, 2001) from a pool of spectral, structural, and texture features. We settled on the *normalized difference vegetation index* (NDVI), *normalized digital surface model* (NDSM) and the pixelwise red, green and near infrared spectral components.

Table 1. Dataset properties. *GSD*: ground sampling distance. R/G/B/I: red / green / blue / near infrared band; patches: number of patches per data set; features / classes: numbers of features used / classes discerned in classification

Dataset	GSD	Channels	Patches	Features	Classes
Vaihingen	8 cm	RGI	15	5	4
Potsdam	8 cm	RGBI	23	5	4
3CITYDS	20 cm	RGBI	27	5	4

In this section, we present an experimental evaluation for two different data acquisition settings. The first, *domain selection*, corresponds to a setting in which only one new target image needs to be classified while large quantities of labeled images are already available from earlier surveys. For the *domain ranking* setting, we assume that a large amount of target images has to be classified and that initially no training data is available, so that domain ranking is applied to determine which images should be labeled to serve as source domains. In all experiments, the evaluation is based on metrics derived from the overall accuracy (OA), i.e., the percentage of correctly classified pixels when comparing the classification results to a reference.

Domain Selection

A successful source selection should be able to find related sources and reduce the expected classification error. The evaluation consists of two parts. First, we analyze our proposed multi-source selection method. Our method is applied to each patch (=T) to synthesize a source $\overline{S}$ using all remaining patches of the dataset as candidate sources. For the *domain selection* setting, we assume that these candidate sources are fully labeled. We examine several source selection strategies. *Single source selection* selects only one source domain that has the lowest domain distance to the target domain while *multi-source selection* utilizes labeled samples from all source domains using source weights as previously described. We examine both strategies in combination with both domain distances $d_{\{SDA,UDA\}}$ and their asymmetric variants $d_{\{A\text{-}SDA,A\text{-}UDA\}}$.

We compare these methods to two simple reference methods: *Random Source* and *All Sources*. *Random Source* selects a single source randomly from all candidate sources. *All Sources*, on the other hand, uses all sources and assigns them uniform source weights. In the first set of experiments, we are mainly interested in the performance of the synthesized source on the target task, so that classification is performed using multi-class logistic regression without DA, but using the source weights π_{S^s} to weight the samples.

In our second experiment, we enable the DA extension for our classifier, applying it to a synthesized source $\overline{S}$ generated by our unsupervised asymmetric multi-source selection algorithm using only the 1 to 3 sources featuring the largest source weights.

Source selection and DA are applied using pixels on a regular grid of size 10 px to 30 px to reduce spatial dependency; the grid size was adapted to the GSD and the patch size of the individual datasets, thus using only about 0.25 percent of the data in these processes (while using all data for evaluation). For the source selection, we selected about 80 percent of these pixels per patch for each bootstrap run. For the logistic regression classifier, we applied a polynomial expansion of degree 2. The entire set of parameters used for DA is given in Table 2, whereas Table 3 shows the parameters used for source selection. The DA parameters were tuned empirically on a small random subset of patches across all datasets. The same parameter values were used for all datasets without further tuning. The source selection parameters are non-critical and were set to achieve a good tradeoff between speed and performance. As source selection has some random components, each experiment is repeated ten times, and we report average quality indices.

Table 2. Parameters used for the DA method previously described. σ_0, σ_{DA}: Weights for the gaussian priors for regularization used for training the initial classifier and in the DA process, respectively. ρ_E, ρ_A: number of samples per class for transfer and elimination. KNN: number of neighbors in the KNN analysis for deciding which target samples to include for training. h: parameter of the weight function $g^{\frac{i}{TD},h}$ (Paul *et al.*, 2016). $i^{\max}$, $g_{P,S}^{max}$, $g_{P,T}^{max}$: parameters of the weight function in Equation 4, in case of $g^{\max}$ for source and target domain, respectively.

σ_0	σ_{DA}	ρ_E	ρ_A	KNN	h	i^{max}	$g_{P,S}^{max}$	$g_{P,T}^{max}$
35	15	30	30	19	0.7	200	1.5	0.9

Table 3. Parameters of multi-source selection. MaxIter GSS: Maximum number of iterations of Golden-Section-Search. INN λ: parameter of the weight function in equation 22. INN β: threshold for the sum of weights for generating a synthetic source domain. Bootstrap runs / size: number of bootstrap runs for synthetic source generation and number of samples used in each run, respectively. N_{max}: number of samples used to determine the asymmetric domain distance.

MaxIter GSS	INN λ	INN β	Bootstrap Runs	Bootstrap Size	N_{max}
10	0.5	0.9	10	5000	60

Domain Ranking

1. Source: Extract from the geospatial data of the Lower Saxony surveying and cadastral administration, ©2013 LGLN

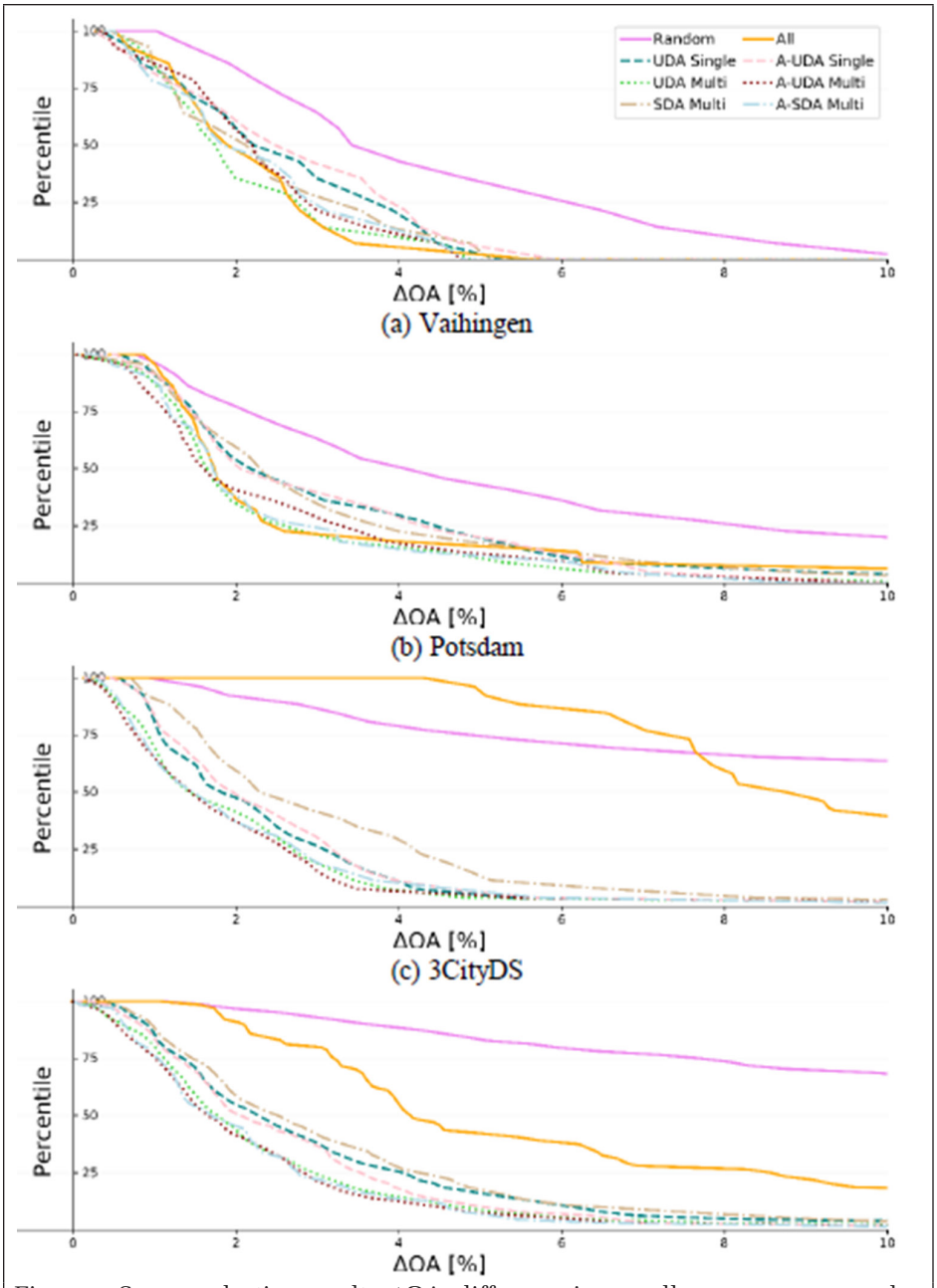


Figure 5. Source selection results. ΔOA: difference in overall accuracy compared to a classifier based on target training data. Percentile: the percentage of patches in the data set for which ΔOA is smaller than the value on the abscissa. Example (Vaihingen, *All Sources*): for 25% of the target patches the loss in OA is larger than 3% (ΔOA > 3%).

Table 4. Source selection results for different variants of the algorithm as previously explained. Mean ΔOA: Average loss in overall accuracy when compared to a classifier based on target training data in 10 test runs (lower is better). STDEV: standard deviation of ΔOA over 10 test runs.

		Random	All	UDA Single	A-UDA Single	UDA Multi	A-UDA Multi	SDA Multi	A-SDA Multi
Vaihingen	Mean ΔOA	4.4	2.2	2.5	2.7	2.1	2.3	2.5	2.3
	Stdev	2.9	1.3	1.5	1.6	1.3	1.3	1.4	1.3
Potsdam	Mean ΔOA	6.2	3.1	3.4	3.1	2.5	2.6	3.3	2.5
	Stdev	5.9	3.5	3.3	2.4	2.3	2.3	3.0	2.1
3CITYDS	Mean ΔOA	26.6	10.6	2.6	2.7	2.3	2.2	3.4	2.3
	Stdev	22.5	5.0	2.8	2.8	2.8	2.7	3.3	2.6
Combined	Mean ΔOA	20.5	7.5	3.2	2.8	2.5	2.3	3.3	2.4
	Stdev	15.6	7.4	2.9	2.6	2.5	2.3	2.8	2.2

For this experiment, we evaluate our proposed domain ranking algorithm. The goal is to achieve a high overall accuracy while only using sources from a small set of candidates, thus reducing the work related to manually labeling these sources. Therefore, we only use the most informative domains as source candidates as defined by the domain ranking produced by the *kernel herding* algorithm. Previous experiments have shown that single source selection with our new asymmetric domain distance ($d_{\text{A-UDA}}$) is competitive with our best multi-source method while also being much faster to compute. For this reason, we ran the *domain ranking* experiments using this source selection method only. To give a context to our results, we also provide an upper and lower bound of the average overall accuracy for the datasets. When only a single labeled source was used, the upper bound was determined by testing all patches as sources, selecting the source that maximized the average overall accuracy over the entire dataset. The bound for larger sets of sources was estimated in a greedy manner by iteratively adding source candidates using the same criterion. The lower bound was generated similarly by minimizing the average overall accuracy.

Results and Discussion

Domain Selection

Figure 5 and Table 4 show the evaluation of source selection without using DA. The evaluation is based on $\Delta OA = OA_{TT} - OA_{ST}$, where OA_{TT} is the overall accuracy achieved on the target dataset when training the classifier on a labeled target dataset and OA_{ST} is the overall accuracy on the target dataset when training on a synthesized source. Thus, ΔOA directly shows how much performance is lost by not having access to class labels in the target domain, and it should be as small as possible. We present percentile plots and the average ΔOA as well as the standard deviation (STDEV) of ΔOA over 10 test runs for each dataset separately. The percentile plots show the cumulative distribution of ΔOA over all patches in a dataset. Generally, we strive to achieve large losses (right side on the percentile plots) for only a small number of patches in a dataset (bottom of the percentile plots). The results do not exhibit too many surprises. With all datasets, random selection is clearly inferior to all other tested methods. Furthermore, using multiple weighted sources usually outperforms single source selection. Our asymmetric MMD generally performs similarly to their symmetric versions. Yet, while the MMD is evaluated on the entire

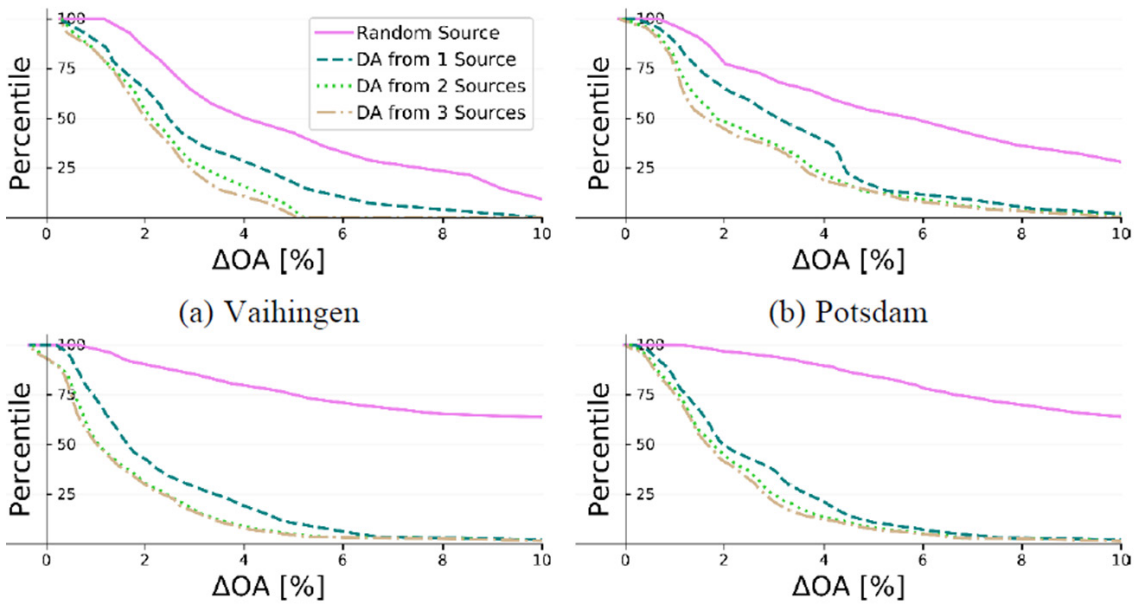


Figure 6. Multi-source domain adaptation results. ΔOA: difference in overall accuracy compared to a classifier based on target training data. For the interpretation of the figures, cf. Figure 5.

training sets, the AMMD only ever has access to N_{max} samples of each source training set. In contrast to the experiments in (Vogt *et al.*, 2017), the supervised domain distances generally perform worse than their unsupervised variants. The core idea is that the d_{SDA} adds a bias to prefer sources that have a larger margin, and therefore also a simpler decision boundary. It appears that this bias might not always be desirable and its application should depend both on the feature space and on the classification method. Surprisingly the AMMD seems to be less affected. We currently do not have an explanation for this observation. For the Vaihingen dataset, the tested methods result in very similar results, which is a behavior different from the one for the other datasets. As most patches in the Vaihingen dataset have a very similar appearance and class distribution, the gains from using TL methods should be expected to be small. The 3CITYDS and Combined datasets, on the other hand, present a more difficult challenge due to the pronounced inhomogeneity between patches. While both naive source selection strategies, *Random Source* and *All Sources*, perform particularly bad here, our proposed multi-source selection methods manage to achieve stable performance (≲2.5%) across all datasets.

Figure 6 and Table 5 show the DA results using a random source and the 1 to 3 best sources according to unsupervised multi-source selection using our asymmetric MMD metric. Again, the evaluation is based on ΔOA as described earlier in this section. In addition, we compared the OA on the target data with and without enabling the DA extension in logistic regression.

We report $\Delta DA = OA_{ST}^{DA} - OA_{ST}$, where OA_{ST}^{DA} is the overall accuracy on the target dataset when training on a synthesized source after domain adaptation. ΔDA can be understood as the mean difference in OA due to enabling DA over all patches of a dataset, where positive ΔDA represents a positive transfer. The test shows that using multiple sources always improves the prospects of DA (indicated by $\Delta DA > 0$), but this effect also seems to diminish quickly when a larger number of sources is used. Compared to the results in (Vogt *et al.*, 2017) the gains of using DA seem to be reduced for more complex feature spaces. It can be observed that DA still shows the greatest benefits for complex and inhomogeneous datasets, like the 3CITYDS or *Combined* datasets. Our initial working hypothesis was that applying instance-transfer based DA to a related source should improve the expected gains, with the goal to achieve positive transfer in most target domains. Our experiments have shown that while selecting a related source is a necessary condition to this end, it does not appear to be sufficient alone.

Table 5. Multi-source domain adaptation results. Mean ΔOA: the average loss in overall accuracy (OA) after DA when compared to a classifier based on target training data (lower is better); the average of 10 test runs is reported. Stdev: standard deviaton of ΔOA over 10 test runs. ΔDA: the improvement in OA when enabling DA (higher is better). DA1-3 applies domain adaptation to the best one to three sources based on our unsupervised asymmetric multi-source selection.

	Random	DA1	DA2	DA3
Mean ΔOA	5.3	3.2	2.4	2.2
Stdev	3.9	2.6	1.5	1.4
ΔDA	-0.9	-0.5	0.0	0.1

(a) Vaihingen

	Random	DA1	DA2	DA3
Mean ΔOA	8.0	3.5	2.8	2.6
Stdev	7.7	2.7	2.5	2.4
ΔDA	-1.8	-0.4	-0.1	-0.1

(b) Potsdam

	Random	DA1	DA2	DA3
Mean ΔOA	26.3	2.6	2.0	1.9
Stdev	22.1	2.8	2.8	2.8
ΔDA	0.4	-0.1	0.2	0.3

(c) 3CityDS

	Random	DA1	DA2	DA3
Mean ΔOA	19.6	2.8	2.4	2.3
Stdev	15.4	2.6	2.4	2.3
ΔDA	0.8	-0.3	0.0	0.1

(d) Combined

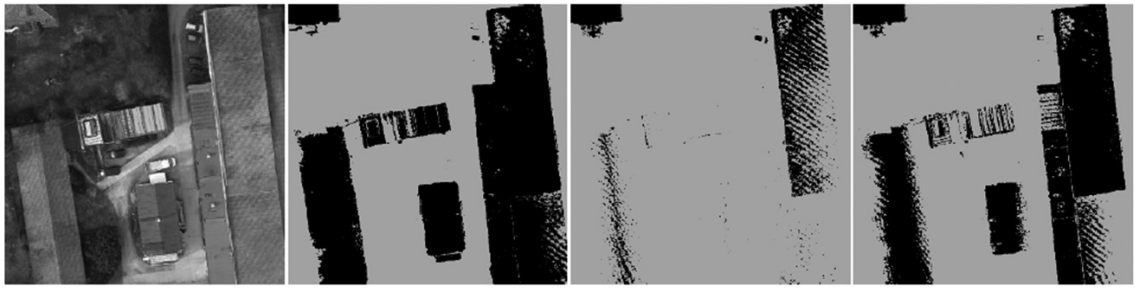
Despite the modest improvements in overall accuracy, DA may still be worthwhile for some applications. Figure 7 shows an example for the class *building* from our DA experiments using the *Combined* dataset. The figure shows that the synthesized source sometimes failed to reproduce low buildings with flat roofs; obviously, even in the synthesized source the DSM heights were not representative for such buildings in these cases. These buildings may be recovered using DA, as seen in Figure 7d. The overall pixel count covered by such objects remains small compared the patch size, which explains their low impact on the measured ΔDA values.

Domain Ranking

Figure 8 shows the results of our *domain ranking* experiments. The diagrams plot the average OA for a dataset when applying source selection as a function of the number N_I of source domains that are assumed to provide training data. The order in which the domains are considered for labeling and, thus, to be included in the set of available source domains, is the one predicted by our *domain ranking* procedure. It can be easily seen that our proposed method is capable of selecting the most important sources with a high degree of certainty. The results of our kernel herding approach follow the theoretical optimum very closely on all datasets. For the *Vaihingen, Potsdam* and 3CITYDS datasets, less than five of the patches would have to be labeled manually to achieve results closer than 2 percent in OA to a fully labeled dataset. For the *Combined* dataset, this figure can be stated as less than 10 patches. Considering the evaluated datasets, our proposed algorithm would be able to save more than 66 percent to 85 percent in manual labeling cost while only incurring a negligible amount of loss in OA. While the performance for few candidate sources is already quite satisfactory, the plots also show a very slow convergence to the optimum afterwards. It appears that while our proposed kernel matrix does contain enough information to confidently rank the most important domains, it cannot do so for the more uninformative domains. We tested this hypothesis by repeating kernel herding with small random perturbations to K. We notice that the absolute domain ranking quickly becomes unstable after the first few ranks. Yet, for practical applications, we do not expect this to become a significant problem.

We also provide runtime measurements for our single source selection based on the $d_{\text{A-UDA}}$ domain distance. For instance, in the experiments based on the *combined* dataset, computing the source weights for a single target takes 6.6 sec using our GPGPU implementation[2] on a single NVIDIA GTX 1060. Applying *domain ranking* on this dataset therefore takes only about seven minutes. It should be noted that this performance scales linearly with the size of the target training set, the number of source domains and the number of features, yet remains constant with reference to the sizes of the source training sets.

2. Can be made available by the corresponding author on request



(a) Visual (b) Target Classifier (c) Source Selection (d) Domain Adaptation

Figure 7. Example for the classification results for class *building* from the *Combined* dataset. Buildings are printed black. (a) Image (b) Results of a classifier trained on target data (c) Results after multi-source selection without DA using three sources (d) Results with DA.

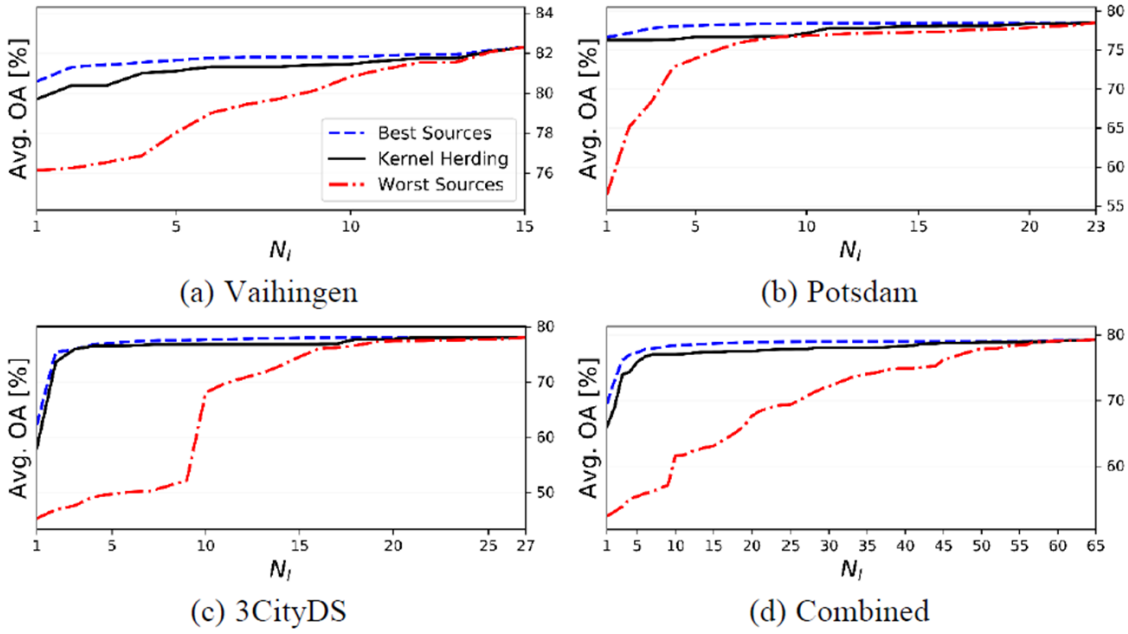


Figure 8. Domain ranking results. Avg. *OA*: average overall accuracy over the dataset for different numbers of candidate source domains (higher is better). N_I: number of source domains that provide labeled training data.

Conclusions

In this work, we presented two domain distance measures based on the MMD and their variants that are able to capture the asymmetric relationship between target and source domains in a supervised learning setting. The supervised domain distances require labeled samples in the source domain, while the unsupervised distances operate without using any labels. We developed a multi-source selection method that synthesizes a related source as a weighted combination of a set of candidate sources, of which only a few may be related to the target. Our fastest method has a linear run-time complexity in regard to the number of candidate sources and the size of the target training set. More importantly, our proposed asymmetric MMD metric has a small memory footprint since it requires less than 100 samples from each source domain and is thus applicable to very large datasets. We also expanded an existing DA method to cope with multiple sources being assigned different weights.

Our experiments show that multi-source selection is consistently able find related sources from a large set of candidate sources. The average loss in classification performance very predictably remains below 2.5 *percent* when compared to a classifier that has full access to labeled samples in the target domain over a variety of datasets. Additionally applying DA achieved a small positive transfer when using the weighted combination of two or more sources selected by our unsupervised procedure. Yet, this gain is quite small and could not be achieved for all datasets. Finally, we examined a scenario where only unlabeled data is available. We applied our source selection method to find the most informative domains. We have shown these informative domains to be good candidates for manual labeling and that an acceptable classification accuracy can be achieved while reducing manual work by up to 85 percent. For our experiments, we have assumed a shared feature space for all domains. In the future, we plan to integrate our source selection method with feature selection and feature extraction approaches, such as deep neural networks (Long *et al.*, 2015). By adaptively finding an optimized feature space in which the target and source domains maximize their similarity, the usage of more complex features should become feasible.

Acknowledgements

This work was supported by the German Science Foundation (DFG) under grants OS 295/4-1 and HE 1822/30-1. The Vaihingen data set was provided by the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF) (Cramer, 2010):
http://www.ifp.uni-stuttgart.de/dgpf/DKEP-Allg.html.

Appendix: Proof for the Relation to Determine N_{max}

Theorem 1: Given a statistically independent sample $X=(x_i)_{i=1}^N$ *from a distribution defined by its cumulative distribution function* $\Pr(x<s)$. *Let* $q=\Pr(x\geq s)$ *be the probability that x is at least as large as a given value s. Also, let* $p=1-\Pr(max_{\in X}x\geq s)$ *be the probability that the largest element in a set X is smaller than s. Then, for a fixed p and q the relationship* $N\geq log_{1-q}p$ *holds.*

Proof. Given

$$q = \Pr(x \geq s) \tag{28}$$

$$p = 1 - \Pr\left(\max_{x \in X} x \geq s\right) = \Pr\left(x < s \,\forall x \in X\right) = (1-q)^N \tag{29}$$

It follows

$$1 - \Pr\left(\max_{x \in X} x \geq s\right) \leq p' \Leftrightarrow (1-q)^N \leq p' \Leftrightarrow N \geq \log_{1-q}(p') \tag{30}$$

References

Acharya, A., E.R. Hruschka, J. Ghosh, and S. Acharyya, 2011. Transfer learning with cluster ensembles, Proceedings of the ICML Workshop on Unsupervised and Transfer Learning, pp. 123–132.

Amini, M.-R., and P. Gallinari, 2002. Semi-supervised logistic regression, Proceedings of the 15th European Conference on Artificial Intelligence, pp. 390–394.

Banerjee, B., F. Bovolo, A. Bhattacharya, L. Bruzzone, S. Chaudhuri, and K. Buddhiraju, 2015. A novel graph-matching-based approach for domain adaptation in classification of remote sensing image pair, IEEE Transactions on Geoscience and Remote Sensing, 53(7): 4045–4062.

Ben-David, S., J. Blitzer, K. Crammer, and F. Pereira, 2007. Analysis of representations for domain adaptation, Advances in Neural Information Processing Systems (NIPS), 19:137–144.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning, First edition, Springer, New York.

Breiman, L., 2001. Random forests, Machine Learning, 45(1):5–32.

Bruzzone, L., and M. Marconcini, 2009. Toward the automatic updating of land-cover maps by a domain adaptation SVM classifier and a circular validation strategy, IEEE Transactions on Geoscience and Remote Sensing, 47(4):1108–1122.

Bruzzone, L., and M. Marconcini, 2010. Domain adaptation problems: A DASVM classification technique and a circular validation strategy, IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(5):770–787.

Chang, M.-W., C.-J. Lin, and R.C. Weng, 2002. Analysis of switching dynamics with competing support vector machines, Proceedings of the International Joint Conference on Neural Networks (IJCNN), Vol. 3:2387–2392.

Chattopadhyay, R., Q. Sun, W. Fan, I. Davidson, S. Panchanathan, and J. Ye, 2012. Multisource domain adaptation and its application to early detection of fatigue, ACM Transactions on Knowledge Discovery from Data, 6(4):18:1–18:26.

Chen, Y., M. Welling, and A. Smola, 2012. Super-samples from kernel herding, arXiv preprint arXiv:1203.3472.

Cheng, L., and S.J. Pan, 2014. Semi-supervised domain adaptation on manifolds, IEEE Transitions on Neural Networks and Learning Systems, 25(12):2240–2249.

Durbha, S., R. King, and N. Younan, 2011. Evaluating transfer learning approaches for image information mining applications, Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 1457–1460.

Eaton, E., M. desJardins, and Y. Lane, 2008. Modeling transfer relationships between learning tasks for improved inductive transfer, Proceedings of the European Conference on Machine Learning (ECML), Springer, pp. 317–332.

Gopalan, R., R. Li, and R. Chellappa, 2011. Domain adaptation for object recognition: An unsupervised approach, Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 999–1006.

Gretton, A., K.M. Borgwardt, M.J. Rasch, B. Schölkopf, and A. Smola, 2012. A kernel two-sample test, Journal of Machine Learning Research, 13(2012):723–773.

Hesterberg, T., D.S. Moore, S. Monaghan, A. Clipson, and R. Epstein, 2003. The Practice of Business Statistics Companion Chapter 18: Bootstrap Methods and Permutation Tests, WH Freeman & Co., New York.

Long, M., Y. Cao, J. Wang, and M.I. Jordan, 2015. Learning transferable features with deep adaptation networks, Proceedings of the 32nd International Conference on Machine Learning (ICML), pp. 97–105.

Matasci, G., D. Tuia, and M. Kanevski, 2012. SVM-based boosting of active learning strategies for efficient domain adaptation, IEEE Journal of Selected Topics in Applied Earth Observation and Remote Sensing 5(5):1335–1343.

Matasci, G., M. Volpi, M. Kanevski,L. Bruzzone, and D. Tuia, 2015. Semisupervised transfer component analysis for domain adaptation in remote sensing image classification, IEEE Transactions on Geoscience and Remote Sensing, 53(7):3550–3564.

Pan, S.J. and Q. Yang, 2010. A survey on transfer learning, IEEE Transactions on Knowledge and Data Engineering, 22(10):1345–1359.

Paul, A., F. Rottensteiner, and C. Heipke, 2016. Iterative re-weighted instance transfer for domain adaptation, ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, III-3:339–346.

Press, W.H., 2007. Numerical Recipes: The Art of Scientific Computing, Third edition, Cambridge University Press, Cambridge, UK.

Schapire, R.E. and Y. Singer, 1999. Improved boosting algorithms using confidence-rated predictions, Machine Learning, 37(3):297–336.

Settles, B., 2010. Active Learning Literature Survey, University of Wisconsin, Madison, Computer Sciences Technical Report 1648.

Shestopaloff, Y.K., 2010. Sums of Exponential Functions and Their New Fundamental Properties, AKVY Press, Toronto, Canada.

Sriperumbudur, B.K., K. Fukumizu, A. Gretton, G.R.G. Lanckriet, and B. Schölkopf, 2009. Kernel choice and classifiability for RKHS embeddings of probability distributions, Advances in Neural Information Processing Systems (NIPS), 22:1750–1758.

Sriperumbudur, B.K., K. Fukumizu, A. Gretton, B. Schölkopf, G.R. Lanckriet, 2012. On the empirical estimation of integral probability metrics, Electronic Journal of Statistics, 6: 1550–1599.

Sugiyama, M., M. Krauledat, and K.-R. Müller, 2007. Covariate shift adaptation by importance weighted cross validation, Journal of Machine Learning Research, 8:985–1005.

Thrun, S. and L. Pratt, 1998. Learning to Learn: Introduction and Overview, (S. Yhrun and L. Pratt, editors), Kluwer Academic Publishers, Boston, Massachusetts, pp. 3–17.

Timofte, R., and L. Van Gool, 2012. Iterative nearest neighbors for classification and dimensionality reduction, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2456–2463.

Tuia, D., J. Munoz-Mari, L. Gomez-Chova, and J. Malo, 2013. Graph matching for adaptation in remote sensing, IEEE Transactions on Geoscience and Remote Sensing, 51(1):329–341.

Tuia, D., E. Pasolli, and W.J. Emery, 2011. Using active learning to adapt remote sensing image classifiers, Remote Sensing of Environment, 115:2232–2242.

Vishwanathan, S., N. Schraudolph, M.W. Schmidt, and K.P. Murphy, 2006. Accelerated training of conditional random fields with stochastic gradient methods, Proceedings of the 23rd International Conference on Machine Learning (ICML), pp. 969–976.

Vogt, K., A. Paul, J. Ostermann, F. Rottensteiner, and C. Heipke, 2017. Boosted unsupervised multisource selection for domain adaptation, ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, IV-1/W1:229–236.

Wegner, J.D., F. Rottensteiner, M. Gerke, and G. Sohn, 2016. The ISPRS 2D Labeling Challenge, URL: http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html (last date accessed: 15 January 2018).

Zadrozny, B., 2004. Learning and evaluating classifiers under sample selection bias, Proceedings of the 21st International Conference on Machine Learning (ICML), pp. 114–121.

Zaremba, W., A. Gretton, and M. Blaschko, 2013. B-test: A non-parametric, low variance kernel two-sample test, Advances in Neural Information Processing Systems (NIPS), Vol. 26, pp. 755–763.

Zhang, Y., X. Hu, and Y. Fang, 2010. Logistic regression for transductive transfer learning from multiple sources, Advanced Data Mining and Applications, Part II, Lecture Notes in Computer Science (L. Cao, J. Zhong, and Y. Feng, editors), Springer, Vol. 6441, pp. 175–182.

WHO'S WHO IN ASPRS

BOARD OF DIRECTORS

BOARD OFFICERS

President
Anne Hillyer
Bonneville Power Administration (USDOE)

President-Elect
Thomas Jordan
University of Georgia

Vice President
Jeff Lovin
Woolpert

Past President
Rebecca A. Morton
GeoWing Mapping, Inc

Treasurer
Stewart Walker

Secretary
Roberta E. Lenczowski

BOARD MEMBERS

Corporate Members Council – 2019
Chair: Joe Cantz
Vice Chair: TBA
www.asprs.org/About-Us/Sustaining-Members-Council.html

Early-Career Professionals Council – 2019
Chair: Jessica Fayne-Kumor
Vice Chair: Amanda Aragon

Region Officers Council – 2019
Chair: Jason S. Smith
Vice Chair: Lorraine Amenda

Student Advisory Council – 2019
Chair: Jeff Pu
Vice Chair: Victoria Scholl
http://www.asprs.org/Students/Student-Advisory-Council.html

Technical Division Directors Council – 2020
Chair: John McCombs
Vice Chair: Bandana Kar

TECHNICAL DIVISION OFFICERS

Geographic Information Systems Division – 2019
Director: Bandana Kar
Assistant Director: Xan Fredericks
www.asprs.org/Divisions/GIS-Division.html

Lidar Division – 2020
Director: Amar K. Nayegandhi
Assistant Director: Joshua Nimetz
www.asprs.org/Divisions/Lidar-Division.html

Photogrammetric Applications Division – 2020
Director: Paul C. Bresnahan
Assistant Director: Kurt Rogers
www.asprs.org/Divisions/Photogrammetric-Applications-Division.html

Primary Data Acquisition Division – 2019
Director: Allen Cook
Assistant Director: Jon Christopherson
www.asprs.org/Divisions/Primary-Data-Aquisition-Division.html

Professional Practice Division – 2020
Director: Jeffrey L. Padgett, C.P.
Assistant Director: Harold W. Rempel, III, CP
www.asprs.org/Divisions/Professional-Practice-Division.html

Remote Sensing Applications Division – 2020
Director: David W. Kreighbaum
Assistant Director: Raechel A. White
www.asprs.org/Divisions/Remote-Sensing-Applications-Division.html

Unmanned Autonomous Systems (UAS) – 2019
Director: Benjamin Vander Jagt
Assistant Director: Megan Miller

REGION PRESIDENTS

Alaska Region
Kirk M. Contrucci, CP
http://www.asprsalaska.org/

Central New York Region
John Boland, C.P.
http://www.esf.edu/asprs/

Columbia River Region
Marcus Glass , C.P.
http://columbia.asprs.org/CRR/

Eastern Great Lakes Region
Gregory N. Lemke, C.P.
http://egl.asprs.org/

Florida Region
Ekaterina Fitos, GISP
http://florida.asprs.org/

Heartland Region
David W. Gwynn, PLS, CP
http://heartland.asprs.org/

Intermountain Region
Sowmya Selvarajan, Ph.D.
http://asprsintermountain.weebly.com/

Mid-South Region
Demetrio Zourarakis
https://www.asprs.org/all-regions/mid-south.html

New England Region
TBA
http://neweng.asprs.org/

North Atlantic Region
Richard W. Carlson, Jr., P.L.S., C.P.
http://natlantic.asprs.org/

Pacific Southwest Region
Lorraine B. Amenda, PLS, CP
https://pswasprs.org/

Potomac Region
James Chris McGlone, CP
http://www.asprspotomac.org/

Puget Sound Region
Michael Rosen
http://puget.asprs.org/

Rocky Mountain Region
Michaela Buenemann
http://www.asprs-rmr.org/

Western Great Lakes Region
Peter Jenkins
http://wgl.asprs.org/

Multitemporal Classification Under Label Noise Based on Outdated Maps

Alina E. Maas, Franz Rottensteiner, Abdalla Alobeid and Christian Heipke

Abstract

Supervised classification of remotely sensed images is a classical method for change detection. The task requires training data in the form of image data with known class labels. If the training labels are acquired from an outdated map, the classifier must cope with errors in the training labels. These errors (label noise) typically occur in clusters in object space, because they are caused by land cover changes over time. In this paper we adapt a label noise tolerant training technique for classification, so that the fact that changes affect larger clusters of pixels is considered. We also integrate the existing map into an iterative classification procedure to act as a priori in regions which are likely to contain changes. Additionally we expand the model for multitemporal data, making it applicable for time series. Our experiments are based on four test areas, including a multitemporal example. Our results show that this method helps to distinguish between real changes over time and false detections caused by misclassification and thus improve the accuracy of the classification results.

Introduction

The updating of topographic databases (referred to as *maps* for brevity) is typically based on a classification of current remote sensing imagery. Comparing the results to the map, areas of change can be detected and the map can be updated accordingly. Supervised classification is commonly used for that purpose, requiring representative training data that are typically generated in a time-consuming manual process. The latter could be avoided by using the existing map to derive the class labels of the training samples. As the map may be outdated, classifiers using the class labels derived from the map for training must take into account the fact that some of these labels will be wrong. Nevertheless, changes typically only affect a relatively small part of a scene, so that one can assume the majority of the training data to be correct.

From the point of view of training the classifier, changes will correspond to errors in the training labels if the outdated map is used to obtain the training samples. In machine learning, such errors in the class labels of training data are referred to as *label noise* (Frénay and Verleysen, 2014). In remote sensing, the problem has mostly been dealt with by data cleansing, i.e., by detecting and eliminating wrong training samples, e.g., Radoux *et al.* (2014). An alternative is to use probabilistic methods for training under label noise which also estimate the parameters of a noise model. An example for such an approach is the label noise tolerant logistic regression (Bootkrajang and Kabán, 2012), which has been applied successfully in the context of remote sensing in (Maas *et al.*, 2016). However, the underlying noise model of that technique assumes wrong labels to occur at random positions in the image. This is not a very realistic model for change detection, where changes typically occur in spatial clusters in object space, e.g., due to the construction of a new building, and may lead to a degradation of the classification performance.

Institute of Photogrammetry and Geoinformation, Leibniz Universität, Hannover, Germany (maas@ipi.uni-hannover.de)

Using the existing map has another potential benefit. As change is usually a rare event, the existing class labels can be seen as providing observations for the prediction of the new class labels. This may be particularly useful in areas where the classifier cannot distinguish the class label by the given features well, e.g., at object borders. The corresponding probabilities for the classes to be correct are related to the probability of observing a wrong label and, thus, to the parameters of a probabilistic noise model (Bootkrajang and Kabán, 2012). However, such an assumption also neglects the fact that changes typically occur in compact clusters. It would typically lead to a strong bias for maintaining the class label of the map, which is desired in areas without changes, but may limit the prospects of detecting real changes.

The parameters of a probabilistic noise model can also be used in a multitemporal setting. The trained parameters describe the change between two epochs, namely the epoch of the map creation and the epoch of recording the current data. If there are remotely sensed data for the first epoch as well, e.g., because the map was created by classifying remotely sensed data, the parameter of the noise model can be used for temporal transitions in multitemporal models like the multitemporal CRF described in (Hoberg *et al.*, 2015).

In this paper, we propose a new supervised classification method that tries to extract as much benefit as possible from the availability of outdated information about the area to be classified, such the existing map and the remotely sensed data of earlier epochs. First, our method uses the class labels from the map for training. This is achieved by expanding the method by Bootkrajang and Kabán (2012) to take into account that changes typically occur in clusters, which we expect to improve the results in scenes with a large amount of change. Second, the class labels of the existing map are included as observations in a classification procedure based on Conditional Random Fields (CRF). We propose an iterative procedure to reduce the impact of the observed class labels in compact areas that are likely to have changed, which we expect to improve the classification results in areas of weak features without affecting the detection of real changes too much. Third, we integrate more epochs into the CRF model to see if the results improve due to use of time series.

To evaluate the inclusion of map information in training and classification we use four datasets with different degrees of changes. One of these datasets also contains several epochs, which is used to evaluate the multitemporal model.

This is an extended version of (Maas *et al.*, 2007). Compared to the original submission, we have expanded the methodology to a multitemporal CRF-based model. We have also expanded the experimental evaluation by adding a new multitemporal dataset of Las Vegas, which, unlike the data in

Photogrammetric Engineering & Remote Sensing
Vol. 84, No. 5, May 2018, pp. 263–277.
0099-1112/18/263–277

doi: 10.14358/PERS.84.5.263

the original paper, contains real changes and, thus, allowed an evaluation of the described procedure with respect to its ability of detecting real changes.

Related Work

The detection of changes, which forms the basis of the updating process of maps, can be based on tree strategies (Jianya *et al.*, 2008). The first group compares the image data of two epochs directly, using features such as band ratios e.g., Subudhi *et al.* (2014). A strong weakness of such methods is the assumption that the appearance of objects remains constant over time, which may not always be correct (Mas, 1999). In addition, the type of change is not determined in most cases, only the occurrence of changes (Lu *et al.*, 2004). The second group compares the results of an independent classification applied to images of both epochs. In such a setting, classification errors directly lead to errors in change detection and thus a sufficient amount of high quality training data are often necessary to receive good classification results (Lu *et al.*, 2004). The most general case is captured by the third group, which integrates all known data simultaneously for multitemporal classification. In a probabilistic context this can be done by Conditional Random Fields, where transition probabilities between epochs are considered, e.g., Hoberg *et al.* (2015). In most cases, these transition probabilities are found empirically, e.g, Hoberg *et al.* (2015), or based on expert knowledge, e.g., Melgani and Serpico (2003). A possibility to extract these probabilities by an expectation maximization algorithm is shown by Bruzzone *et al.* (1999). The algorithm uses two images with the same resolution to extract the joint probability of two epochs. In our case, no image is given for the time of map creation. Additionally the algorithm needs correct training data for both epochs, which we want to avoid.

As no sensor data are assumed to be available for the time of the acquisition of the existing map, we classify the current data and compare the results to the outdated map, which corresponds to the second strategy of change detection. However, we use a multitemporal model for classification if data of more epochs are available, which would correspond to the third strategy.

For the reasons pointed out in the Introduction, a training procedure taking the class labels of the training samples from an existing map must cope with label noise. Frénay and Verleysen (2014) differentiate three types of statistical models for label noise. The *noisy completely at random* (NCAR) model does not consider dependencies between label noise and other variables. In the *noisy at random* (NAR) model, the probability of an error depends on the class label. If the dependencies between labeling errors and the observed data are considered, the model is called *noisy not at random* (NNAR). This would be an appropriate choice in our case to model that label noise typically occurs in clusters in image space. We do not build a NNAR model explicitly, but we use one implicitly by an iterative strategy for reducing the impact of training samples forming clusters of potentially changed pixels. Existing NNAR models tend to analyze the distributions of the training samples in feature space, e.g., assuming label noise to occur more likely near the classification boundaries or in low-density regions (Sarma and Palmer, 2004). Apart from being drawn from another domain than image classification, this is not a model of local dependencies between label noise at neighboring data sites.

Frénay and Verleysen (2014) distinguish three strategies for dealing with label noise. First, classifiers that are robust to a low noise level by design can be used. For example the random forests classifier (Breiman, 2001) is robust to some degree of noise (Pelletier *et al.*, 2017), but still may have difficulties with large amounts of label noise (Maas *et al.*, 2016). The second strategy tries to remove training samples affected by label noise from the training set, e.g., Sun et al. (2007). Such *data cleansing* methods have been criticized for eliminating too many instances (Frénay and Verleysen, 2014). The third option is to use a classifier which is tolerant to label noise. In this context, probabilistic approaches learn the parameters of a noise model along with the classifier in the training process; examples are (Bootkrajang and Kabán, 2012), using logistic regression as the base classifier, and (Li *et al.*, 2007), presenting a method based on the kernel Fisher discriminant. An example for a non-probabilistic approach is the label noise tolerant version of a Support Vector Machine (SVM) (An and Liang, 2013). However, non-probabilistic methods typically do not estimate the parameters of a noise model, e.g., transition probabilities containing the probability for the observed label to be affected by a change (Bootkrajang and Kabán, 2012). Such transition probabilities can be used as temporal transition matrices, linking the observed class labels of the map to class labels at the second epoch (Schistad Solberg *et al.*, 1996). Patrini *et al.* (2017) also use transition probabilities to train deep neural networks, but for different applications, e.g., image retrieval. In the domain of remote sensing, classification under label noise seems to be based on data cleansing in most cases. An example is Radoux *et al.* (2014), where two techniques for eliminating outliers to derive training data from an existing map are presented. The first technique removes training samples near the boundaries of land cover types and the other one removes outliers based on a statistical test, assuming a Gaussian distribution of spectral signatures. Designed for data of 300 m ground sampling distance (GSD), the model assumptions, e.g., Gaussian distributions, cannot be used directly for high resolution images. A similar method was used for map updating in (Radoux and Defourny, 2010), using Kernel density estimation for deriving probability densities. Another data cleansing method is reported in Jia *et al.*, (2014). Similarly to the method proposed in this paper, all pixels from an existing map are used for training and the resulting label image is compared to the existing map to detect changes. However, no parameters of a model for label noise are estimated in the training process. This is also true for the data cleansing method based on SVM by Büschenfeld (2013), who eliminates training samples that are assigned to another class than indicated by the given map or that show a high uncertainty. Label noise tolerant training using maps for generating training labels was applied by Mnih and Hinton (2012). They propose two loss functions tolerant to label noise to train a deep neural network, but their method only deals with binary classification problems. Bruzzone and Persello (2009) include information of the pixels in the neighborhood of the training samples in the learning process to achieve robustness to label noise in a context-sensitive semi-supervised SVM. Although the authors argue that such a strategy can be used to integrate existing maps for training, this is not shown explicitly.

In Maas *et al.* (2016), label noise tolerant logistic regression (Bootkrajang and Kabán, 2012) was applied to use an existing map for training, integrating it into a CRF for context-based classification. The experiments showed that the method is tolerant to a large amount of label noise if it is randomly spread over the image, as would be expected for a method based on a NAR model. However, experiments with more realistic changes were only shown with a small percentage of wrong training labels, and the class labels from the existing map were not used in the classification process. The latter was done by Schistad Solberg *et al.* (1996), who applied a temporal model based on transition probabilities to include an

outdated land cover map in multitemporal classification, but no local dependencies between changes were considered.

In this paper we build on the method described in Maas *et al.* (2016), but we expand it by considering the fact that changes occur in clusters. Label noise logistic regression (Bootkrajang and Kabán, 2012) is applied in an iterative procedure in which the impact of training samples in areas of potential change is reduced, while these samples are not completely eliminated. To consider local context, the resultant classifier is integrated in a CRF, in which we consider the original class labels as additional observations. In contrast to Schistad Solberg *et al.* (1996), the influence of these observations may change in the course of an iterative process if a pixel is situated in a large cluster of potentially changed pixels, so that temporal oversmoothing (Hoberg *et al.*, 2015) can be avoided. If data from more than one epoch are available we use a multitemporal CRF as in Hoberg *et al.* (2015), but with three main differences. First, the outdated map or classification results of prior epochs are used for training, instead of manually labeled data. This can be done due to the label noise tolerant training procedure. Second, unlike Hoberg et al. (2015) we integrate the labels of an old map in the CRF model as mentioned before. Third, the results from the label noise tolerant training are used to determine the temporal transition probabilities, instead of fixing them to values found empirically. Similar to the integration of the map information these transition probabilities may change if a pixel is likely to be inside a large area of potential change.

Our method can be seen as a combination of "soft" data cleansing (because samples are not eliminated completely) with a probabilistic noise model for including the observed labels from the map, including spatial and temporal context. Thus, we expect it to be able to cope with a larger amount of real change than Maas et al. (2016).

Label Noise Tolerant Change Detection

We assume remotely sensed data from at least one epoch and an existing but outdated raster map to be available in the same coordinate system; note that for the sake of clarity we introduce time indices only in sections dealing with the multitemporal setting. The data of an epoch consist of N pixels, each pixel n represented by a feature vector $\mathbf{x}_n = [x_n^1, \ldots, x_n^F]$ of dimension F, calculated from the imagery, and an observed class label $\tilde{C}_n \in \mathbb{C} = \{C^1, \ldots, C^K\}$ from the existing map. $\mathbb{C}$ denotes the set of classes and K is the total number of classes. As the database may be outdated, the observed labels may differ from the unknown true labels $C_n \in \mathbb{C}$. Collecting the observed and the unknown class labels in two vectors $\tilde{\boldsymbol{C}} = (\tilde{C}_1, \ldots, \tilde{C}_N)^T$ and $\boldsymbol{C} = (C_1, \ldots, C_N)^T$, respectively, and denoting the observed image data by $\mathbf{x}$, it is our goal to find the optimal configuration of class labels $\boldsymbol{C}$ by maximizing the joint posterior $P(\boldsymbol{C} \mid \boldsymbol{x}, \tilde{\boldsymbol{C}})$ of the unknowns given the observations. In this process, we use the class labels of the outdated map for deriving the class labels of the training samples. We start by outlining our modified version of the training procedure for logistic regression by Bootkrajang and Kabán (2012). In the next subsection, we show how logistic regression is integrated into a CRF (Kumar and Hebert, 2006) together with a model for considering the existing class labels $\tilde{\boldsymbol{C}}$ as observations. The next subsection describes the new iterative procedure for training and inference; these are the sections in which images from one epoch only are assumed to be available. In the last subsection, we expand the CRF model to integrate remotely sensed data from multiple epochs.

Label Noise Robust Logistic Regression

Classification is based on logistic regression, a discriminative probabilistic classifier that directly models the posterior probability $p(C_n \mid \mathbf{x}_n)$ of a class label C_n given the feature vector $\mathbf{x}_n$. A feature space transformation $\boldsymbol{\Phi}(\mathbf{x}_n)$ may be applied to achieve non-linear decision boundaries in the original feature space. In the multiclass case, the posterior is modeled by Bishop (2006):

$$p\left(C_n = C^k \mid \mathbf{x}_n, \mathbf{w}\right) = \frac{\exp\left(\mathbf{w}_k^T \cdot \boldsymbol{\Phi}(\mathbf{x}_n)\right)}{\sum_{j=1}^{K} \exp\left(\mathbf{w}_j^T \cdot \boldsymbol{\Phi}(\mathbf{x}_n)\right)} \tag{1}$$

where $\mathbf{w}_k$ is a vector of parameters for a particular class C^k. As the sum of the posterior over all classes has to be 1, these parameter vectors are not independent, so that $\mathbf{w}_1$ is set to $\mathbf{0}$; the other vectors are collected in a joint parameter vector $\mathbf{w}$.

In our case, each training sample consists of a feature vector $\mathbf{x}_n$ and the observed label $\tilde{C}_n$. In order to consider this fact in training, Bootkrajang and Kabán (2012) model the probability $p(\tilde{C}_n \mid \mathbf{x}_n)$ as the marginal distribution of the observed labels $\tilde{C}_n$ over all values the unknown class labels C_n may take:

$$p\left(\tilde{C}_n = C^k \mid \mathbf{x}_n, \mathbf{w}\right) = \sum_{a=1}^{K} p\left(\tilde{C} = C^k \mid C = C^a\right) p\left(C_n = C^a \mid \mathbf{x}_n, \mathbf{w}\right) \tag{2}$$

In Equation 2, $p(\tilde{C}=C^k \mid C= C^a)$ is the probability for a specific type of label noise affecting the two classes C^a and C^k. These *transition probabilities* for all class configurations form the K × K *transition matrix* $\boldsymbol{\Gamma}$ with $\boldsymbol{\Gamma}(a,k) = \gamma_{ak} = p(\tilde{C}=C^k \mid C= C^a)$. The transition matrix $\boldsymbol{\Gamma}$ contains the parameters of a NAR model which are estimated along with the parameters $\mathbf{w}$ in Equation 1. Because this kind of model is unrealistic to describe changes in land cover, we introduce a weight $g_n \in (0 \ldots 1]$ for every sample n to control its influence in the training process. In the beginning, these weights are all set to 1; the last subsection describes how they are changed iteratively to consider the assumption that changes occur in local spatial clusters. To determine the unknown parameters $\mathbf{w}$ and $\boldsymbol{\Gamma}$, we apply maximum likelihood estimation of the unknown parameters with a Gaussian prior over $\mathbf{w}$ for regularization. Taking the negative logarithms of the involved probabilities, this results in the minimization of the following target function:

$$E(\mathbf{w}, \boldsymbol{\Gamma}) = -\sum_{n=1}^{N} \left(g_n \cdot \prod_{k=1}^{K} t_{nk} \ln\left(S_{nk}\right) \right) + \frac{\mathbf{w}^T \mathbf{w}}{2\sigma^2} \tag{3}$$

In Equation 3, t_{nk} is an indicator variable taking the value 1 if $\tilde{C}_n = C^k$ and 0 otherwise, $S_{nk} = p(\tilde{C}_n=C^k \mid \mathbf{x}_n, \mathbf{w})$ as defined in Equation 2, and the right-most term corresponds to the logarithm of a Gaussian prior with zero mean and covariance $\sigma \cdot \mathbf{I}$, where $\mathbf{I}$ is a unit matrix. We use the Newton-Raphson method (Bishop, 2006) for minimizing $E(\mathbf{w}, \boldsymbol{\Gamma})$. In each iteration τ, the parameter vector $\mathbf{w}^\tau$ is determined from $\mathbf{w}^{\tau-1}$ according to $\mathbf{w}^\tau = \mathbf{w}^{\tau-1} - \mathbf{H}^{-1} \nabla E$, where $\nabla E = [\nabla_{w_2} E^T, \ldots, \nabla_{w_K} E^T]^T$ is the gradient of $E(\mathbf{w}, \boldsymbol{\Gamma})$:

$$\nabla_{w_j} E = \sum_{n=1}^{N} \left(g_n \cdot \left(f_{nj} - \bar{t}_{nj} \right) \boldsymbol{\Phi}(\mathbf{x}_n) \right) + \frac{1}{\sigma^2} \mathbf{w}. \tag{4}$$

In Equation 4 we use the shorthand $f_{nj} = p(C_n = C^j \mid \mathbf{x}_n, \mathbf{w})$ for the posterior in Equation 1, and $\bar{t}_{nj} = f_{nj} \sum_{k=1}^{K} \left(\gamma_{jk} \frac{t_{nk}}{S_{nk}} \right)$. The

Hessian matrix $\mathbf{H}$ consists of $(K-1)\times(K-1)$ blocks $\boldsymbol{H}_{ij} = \nabla_{w_i}\nabla_{w_j}E$:

$$\nabla_{w_i}\nabla_{w_j}E = \sum_{n=1}^{N}\left(g_n\left(f_{ni}f_{nj}\xi + I_{ij}\left(f_{nj}-\bar{t}_{nj}\right)\right)\boldsymbol{\Phi}(\mathbf{x}_n)\boldsymbol{\Phi}(\mathbf{x}_n)^T\right) + \frac{\delta(i=j)}{\sigma^2}\mathbf{I}. \quad (5)$$

In Equation 5, $\xi = \sum_{k=1}^{K}\left(\gamma_{jk}\gamma_{ik}\frac{t_{nk}}{S_{nk}^2}\right)$, $\mathbf{I}$ is a unit matrix with elements I_{ij}, and $\sigma(i=j)$ is the Kronecker delta function delivering a value of 1 if the argument is true and 0 otherwise. Optimizing for the unknown weights requires knowledge about the transition matrix $\boldsymbol{\Gamma}$, which, however, is unknown. Bootkrajang and Kabán (2012) propose an iterative procedure similar to expectation maximization (EM). Starting from coarse initial values for $\boldsymbol{\Gamma}$, the parameters $\mathbf{w}$ of the classifier are updated as just described. Using these weights, the transition matrix $\boldsymbol{\Gamma}$ is updated afterwards, expanding the updating step presented in Bootkrajang and Kabán (2012) by the weights g_n:

$$\gamma^{\tau} = \frac{1}{c}\gamma_{jk}^{\tau-1}\sum_{n=1}^{N}\left(g_n t_{nk}\frac{f_{nj}}{S_{nk}^{\tau-1}}\right). \quad (6)$$

In Equation 6, $c = \sum_{l}^{K}\left(\gamma_{jl}^{\tau-1}\sum_{n=1}^{N}\left(g_n t_{nl}\frac{f_{nj}}{S_{nl}^{\tau-1}}\right)\right)$ and $S_{nk}^{\tau-1} = \sum_{j=1}^{K}\left(\gamma_{jk}^{\tau-1}f_{nj}\right)$. This alternating update of the parameters $\mathbf{w}$ and $\boldsymbol{\Gamma}$ is repeated until a termination criterion is reached. The estimated parameters $\mathbf{w}$ are related to a classifier delivering the posterior for the unknown current labels C_n, not the noisy labels $\tilde{C}_n$. Note that training with equal weights $g_n = 1$ was already used in Maas *et al.* (2016). In this paper, this is just the case in the beginning of the training procedure. It also has to be noted that the transition matrix $\boldsymbol{\Gamma}$ only represents the transition between the old database and the current labels in the initial training step in which equal weights g_n are used. If the weights of training samples in large clusters of potential changes are low, the majority of the samples affected by label noise will have a low impact on the result, so that $\boldsymbol{\Gamma}$ only represents residual label noise of small local extents for which the NAR model is a sufficiently good approximation.

CRF Considering the Existing Map

CRFs are graphical models consisting of nodes and edges that can be used to consider local context in a probabilistic classification framework (Kumar and Hebert, 2006). The nodes of the underlying graph represent random variables, whereas the edges connect pairs of nodes and represent their statistical dependencies. Here, the unknown nodes correspond to the current labels C_n of all pixels n, and the edges are defined on the basis of a 4-neighborhood on the image grid. As described above, the observed variables are the image data $\mathbf{x}$ and, different from (Kumar and Hebert, 2006), the observed class labels $\tilde{\boldsymbol{C}}$ (cf. Figure 1 for the structure of the graphical model). The joint posterior $P(\boldsymbol{C}\mid\mathbf{x},\tilde{\boldsymbol{C}})$ of the unknowns given the observations is modeled by:

$$P\left(\boldsymbol{C}\mid\mathbf{x},\tilde{\boldsymbol{C}}\right) = \frac{1}{Z}\exp\left[\sum_{n}\left(A_x\left(C_n,\mathbf{x}\right)+A_T\left(C_n,\tilde{C}_n\right)\right)+\sum_{n,m\in\epsilon}\left(I\left(C_n,C_m,\mathbf{x}\right)\right)\right], \quad (7)$$

were Z is a normalization constant and is the set of edges in the graph. The association potential $A_x(C_n, \mathbf{x})$ connects the unknown label C_n of pixel n with the image data $\mathbf{x}$. Its dependency from the entire input image $\mathbf{x}$ is considered by using site-wise feature vectors $\mathbf{x}_n(\mathbf{x})$, which may be functions of larger image regions. Any discriminative classifier can be used to model this potential (Kumar and Hebert, 2006); here, it is based on the posterior $p(C_n\mid\mathbf{x}_n)$ of logistic regression according to Equation 1:

$$A_x\left(C_n,\mathbf{x}\right) = \ln\left(p\left(C_n\mid\mathbf{x}_n\right)\right) \quad (8)$$

The interaction potential $I(C_n, C_m, \mathbf{x})$ describes the statistical dependencies between a pair of neighboring labels C_n and C_m. In this paper, the contrast-sensitive Potts model is used for that purpose, which results in a data-dependent smoothing of the resultant label image (Boykov *et al.*, 2001):

$$I\left(C_n,C_m,\mathbf{x}\right) = \delta\left(C_n,C_m\right)\cdot\beta_0\left(\beta_1+\left(1-\beta_1\right)\cdot e^{\left(-\frac{\Delta\mathbf{x}^2}{2\sigma_D^2}\right)}\right) \quad (9)$$

where the parameters β_0 and β_1 describe the overall degree of smoothing and the impact of the data-dependent term, respectively, σ_D is the average squared distance between neighboring feature vectors, $\Delta\mathbf{x} = \|\mathbf{x}_n - \mathbf{x}_m\|$ is the distance of two feature vectors $\mathbf{x}_n$ and $\mathbf{x}_m$, and $\delta(C_n, C_m)$ is the Kronecker delta function.

The observed labels are related to the unknown class labels by the temporal association potential $A_T(C_n,\tilde{C}_n)$, derived from the probability of the unknown label given the observed one:

$$A_T\left(C_n,\tilde{C}_n\right) = \theta_n\cdot\ln\left(p\left(C_n = C^b\mid\tilde{C}_n = C^a\right)\right). \quad (10)$$

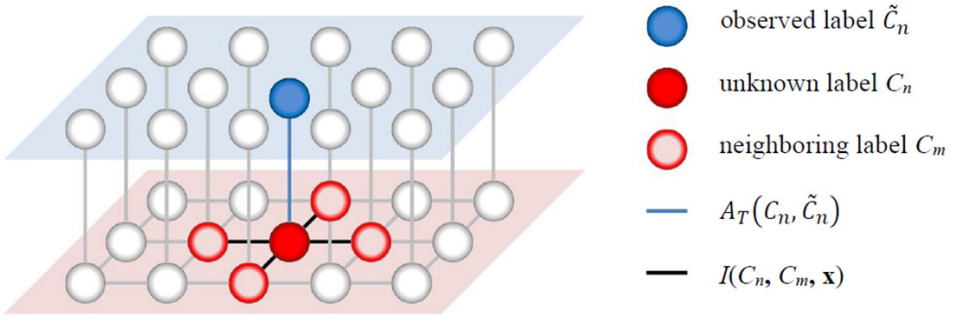


Figure 1. Graph structure of the CRF considering the existing map. The data x are not included for the sake of clarity.

In Equation 8, there is an individual weight $\theta_n \in [0, \ldots, 1]$ for every pixel n. This weight models the influence of the observed label on the classification result of this pixel in inference. As we shall see in the Training and Inference Section, these weights will be adapted in the inference process to reduce the impact of the observed labels for pixels that are very likely to belong to a larger area affected by a change.

Training and Inference

In order to obtain the optimum configuration of the current class labels given the observations by maximizing $P(\boldsymbol{C} \mid \boldsymbol{x}, \tilde{\boldsymbol{C}})$ according to Equation 7, a joint iterative training and inference strategy is applied. After the determination of initial parameters of the association potential and the parameters of the temporal association potentials in an initial training phase, an iterative scheme of classification and re-training is applied in which the weights of pixels in large areas of potential change according to the current classification result are modified to reduce their impact on the results. These steps are described in the subsequent sub-sections.

Initial Training and Classification

In the initial training phase, the observed labels and the data are used for label noise robust training of the logistic regression classifier that serves as the basis for the association potentials of the CRF. For that purpose, the method previousl described is applied, using identical weights $g_n = 1$ for all training samples. This will result in an initial set of parameters $\mathbf{w}$ for the association potentials and a transition matrix $\boldsymbol{\Gamma}$ that contains the transition probabilities $p(\tilde{C}_n = C^u \mid C_n = C^k)$ of the NAR model (Bootkrajang and Kabán, 2012). According to the theorem of Bayes, these probabilities are related to the probabilities $p(C_n = C^k \mid \tilde{C}_n = C^a)$ required for the temporal association potential (Equation 8) by:

$$p\left(C_n = C^k \mid \tilde{C}_n = C^a\right) = \frac{p\left(\tilde{C}_n = C^a \mid C_n = C^k\right) \cdot p\left(C_n = C^k\right)}{p\left(\tilde{C}_n = C^a\right)}. \quad (11)$$

As we have no access to the distribution of the unknown class labels $p(C_n)$, we assume $p(C_n = C_k) \approx p(\tilde{C}_n = C_k)$ to derive the temporal association potential from $\boldsymbol{\Gamma}$. These parameters are kept constant in the subsequent iteration process for the reasons already pointed out: the transition matrix corresponds to the real transition probabilities only in the first iteration (when all training samples have an identical weight $g_n = 1$). The parameters of the interaction potentials (β_0, β_1; cf. Equation 7) are set to values found empirically.

For the determination of the optimal configuration of labels $\boldsymbol{C} = argmax(P(\boldsymbol{C} \mid \boldsymbol{x}, \tilde{\boldsymbol{C}}))$, Loopy Belief Propagation (LBP) is used (Frey and MacKay,1998). In the initial classification, the weights θ_n of the temporal association potentials is set to 0 for all pixels, so that this classification is only based on the current state of the association and the interaction potentials.

Iterative Re-training and Classification

By comparing the current label image to the outdated map, areas of potential change can be detected. This information is used to update the weight g_n of each training sample, and label noise robust training of the logistic regression classifier is repeated, using the updated weights. The way in which the weights are updated will be explained below. Training will result in new values for the parameters $\mathbf{w}$ of the association potentials of the CRF.

Furthermore, the information about potential areas of change is also used to change the weights θ_n of the temporal association potentials as will be explained. Using the updated parameters $\mathbf{w}$ and weights θ_n, another round of inference is carried out, which will lead to an improved classification result. This procedure of updating weights on the basis of the current state of the classification results, re-training and inference is repeated until the proportion of weights that are changed in an iteration is below a threshold or a maximum number of iterations is reached. The procedure is inspired by re-weighting strategies for robust estimation in adjustment theory, e.g., Förstner and Wrobel (2016). The inference results after the last iteration provide the final classification output.

Weights g_n of Training Samples

The weight of a training sample n should be high if that sample is probably not affected by a change, and it should be low otherwise. The weights are initialized by $g_n = 1$ as long as no information about changes is available. After classification, the resulting labels C_n can be compared to the map $\tilde{C}_n$ to generate a binary map B_C of potential changes. However, as indicated in Figure 2b for an aerial image, this binary map will also be affected by classification errors. Thus, we define some simple heuristics to adapt B_C so that it only contains compact clusters of pixels that are very likely to correspond to real changes. As we want our methods to be applicable to a wide variety of images in terms of GSD and land cover, for a specific dataset the user can decide which of these heuristics are to be used. The list of heuristics is as follows:

1. Small objects are likely to correspond to classification errors, and there are characteristic patterns of errors where small objects of a class most frequently belong to another class C^j. To realize such an assumption, a binary map B_k is generated for C^k, containing pixels assigned to C^k as foreground and all other pixels as background. All connected components of foreground pixels in B_k that cover an area smaller than a threshold o_k are removed. After that, the class labels C_n of pixels corresponding to in the original classification result but belonging to the background in B_k are changed to $C_n = C^j$. These labels are compared to the map $\tilde{C}_n$ to generate the binary map B_K of potential changes again.

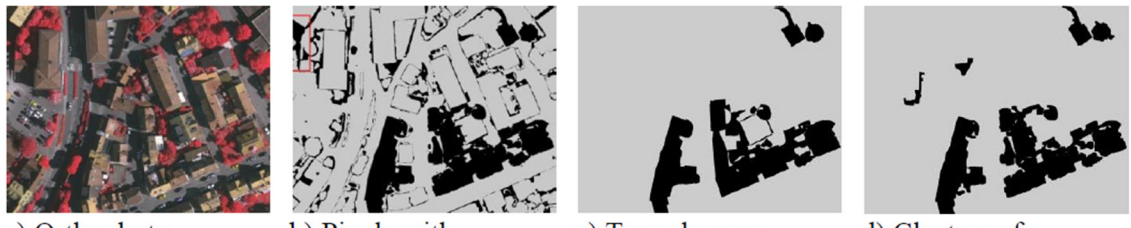


Figure 2. Example for the identification of potential areas of change. Black / gray: changed / unchanged pixels. Red rectangle in (b): a cluster that corresponds to a shadow.

2. Small objects surrounded by a class C^k probably belonging to that class. This is realized by morphological closing of B_k using a structural element of size z_k; again, the labels C_n are changed and a new version of the binary map B_k is generated.
3. Changes occur in clusters. This is considered by removing all connected components of foreground pixels in B_C which cover an area smaller than a threshold u.
4. Classification errors often occur at object boundaries, e.g., because of mixed pixels or because of matching errors if digital surface models (DSM) are used in classification. Thus, a set of connected foreground pixels in B_C forming a line that is thinner than a threshold s is very likely to be caused by classification errors. Such sets are removed by morphological opening using a structural element of size s.
5. In areas affected by cast shadows, the quality of spectral information or of the DSM (if available) is poor and, thus, potential changes as indicated by B_C are very likely to correspond to classification errors. To detect shadow areas, the median and the mean of the image intensity in each cluster cl are compared to the median and the mean of the entire image. If $mean_{cl}<mean_{img}/2$ and $med_{cl}<med_{img}/2$, i.e., if the pixels in the cluster are very dark compared to the image, the pixels belonging to cluster cl are removed from the binary map of potential changes B_C. The remaining foreground pixels in B_C are likely to correspond to real changes (cf. Figure 3d for an example). Of course, this heuristic is only relevant if the GSD is high enough for cast shadows to be visible.

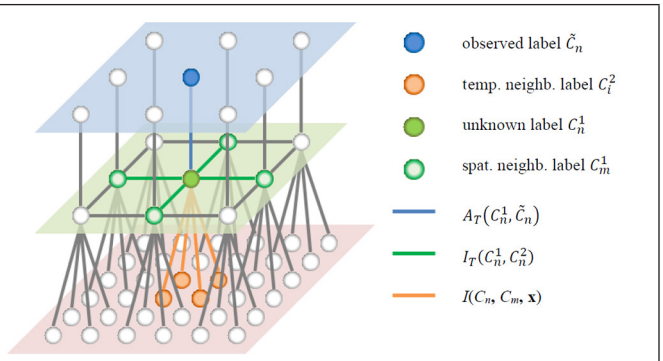


Figure 3. Graph structure of the multitemporal CRF considering the existing map. The data xt are not included for the sake of clarity.

For pixels corresponding to the foreground in the binary map B_C after applying these heuristics, the weights g_n are decreased by a constant c_g, so that in iteration $it+1$, the weight of the corresponding samples is given by $g_n^{it-1} = max(g_n^{it} - c_g, \zeta)$. The minimal weight is set to a small positive constant ζ to avoid numerical problems. The weights of pixels that belong to the background in B_C are updated according to $g_n^{it-1} = min(g_n^{it} + c_g, \zeta)$. As a consequence, the weights of pixels that are likely to correspond to changes will be reduced in each iteration; however, a pixel may regain influence if in a certain iteration its most likely class label is identical to the one from the map, e.g., due to the influence of its neighbors or due to the temporal model.

The weights θ_n of the Temporal Association Potential

The weight θ_n of pixel n regulates the impact of the temporal association potential and, thus, the influence of the outdated map on the resulting label configuration $\mathbf{C}$ (cf. Equation 8). If a pixel n is probably not affected by a change, its weight θ_n should be high, otherwise it should be low. The initial weight for each pixel is 0, because in the beginning we do not want to bias the result to reject potential changes. In the subsequent iterations, the binary map of potential changes B_C used to adapt the weights g_n of the training samples is also used to guide the adaptation of the weights θ_n of the temporal model, because the same assumption w.r.t. the plausibility of a potential change indicated by the current classification results apply. In iteration $it+1$, the temporal association potentials for pixels corresponding to the foreground in B_C will be weighted by $\theta_n^{it-1} = max(\theta_n^{it} - c_\theta, 0)$, where c_θ is a positive constant. The corresponding weights of pixels that belong to the background in B_C are updated by $\theta_n^{it-1} = max(\theta_n^{it} + c_\theta, 1)$.

Multitemporal CRF

If additional images from epochs between the time of creating the map and the epoch of the current image are available, the CRF model has to be expanded. We start with the presentation of the multitemporal CRF model, whereas the second part of this section is dedicated to training and inference in the multitemporal setting.

Multitemporal CRF Model

In the multitemporal setting, we assume the image data to consist of a set of M images with data $\mathbf{x}^t$, where $t \in T$ is the index of the epoch and $T = \{1, \ldots, M\}$ denotes the set of epochs. The images have to be available in the same coordinate system as the existing map (which is supposed to correspond to an epoch prior to $t = 1$). It is our goal to classify the images of all epochs simultaneously, which implies that we have to determine a class label C_n^t for each sample n in each epoch t. Class labels at the same spatial position but corresponding to different epochs are supposed to interact. The resultant graph is shown in Figure 3; note that, similar to Hoberg *et al* (2015), the images at different epochs need not have the same resolution: a pixel in epoch t is connected to all spatially overlapping pixels in epochs $t + 1$ and $t - 1$ if these epochs exist. Consequently, the number of nodes may be different for different epochs. However, the image of the first epoch must have the same resolution as the outdated map. The main difference to Hoberg *et al.* (2015) is that the labels from that map are also included as observations in a way similar to the monotemporal setting. Based on the graph in Figure 3, the joint posterior $P(\boldsymbol{C} \mid \boldsymbol{x}, \tilde{\boldsymbol{C}})$ including the temporal interaction potential is modeled by:

$$P\left(\boldsymbol{C} \mid \mathbf{x}, \tilde{\boldsymbol{C}}\right) = \frac{1}{Z} \exp\left[\sum_t \left(\sum_{n \in v^t} \left(A_x\left(C_n^t, \mathbf{x}^t\right)\right) + \sum_{n,m \in \varepsilon^t} \left(I\left(C_n^t, C_m^t, \mathbf{x}^t\right)\right)\right) + \sum_{n,i \in \kappa} \left(I_T\left(C_n^t, C_i^k\right)\right) + \sum_{n \in v^1} \left(A_T\left(C_n^1, \tilde{C}_n\right)\right) \right] \quad (12)$$

In eq. 12, v^t and ε^t denote the sets of nodes (i.e., pixels) and spatial edges at epoch t, respectively. $A_x(C_n^t, \mathbf{x}^t)$ and $I(C_n^t, C_m^t, \mathbf{x}^t)$ are the corresponding association and (spatial) interaction potentials, respectively; they are defined similarly as in the monotemporal setting, in the case of the association potential using an individual set of parameters $\mathbf{w}^t$ for each epoch. $A_T(C_n^1, \tilde{C}_n)$ is the temporal association potential linking the labels C_n^1 of the first epoch to the observed labels $\tilde{C}_n$ in the same way as previously explained; no map data are assumed to be available for the other epochs. The set of temporal edges is denoted by κ; the corresponding temporal interaction potential $I_T(C_n^t, C_i^k)$ links the labels C_n^t and C_i^k of spatially overlapping pixels in neighboring epochs t and k.

The temporal interaction potential is related to the transition probability, i.e., to the probability of the label C_n^t of time t given the label C_i^k of time k:

$$I_T\left(C_n^t, C_i^k\right) = \vartheta_{n,i}^{t,k} \cdot \ln\left(p\left(C_n^t = C^b \mid C_i^k = C^a\right)\right). \quad (13)$$

Similar to Equation 10, an individual weight $\vartheta_{ni}^{t,k}$ is assigned to each term to model the influence of that term on the results. The definition of these transition probabilities $p(C_n^t = C^b \mid C_i^k = C^a)$ and weights $\vartheta_{ni}^{t,k}$ is explained in the Training and Inference section.

Training and Inference

To obtain the maximal joint posterior $P(\boldsymbol{C} \mid \boldsymbol{x}, \widetilde{\boldsymbol{C}})$ (Equation 12), the multitemporal CRF model must be trained. The parameters to be determined are the weight vectors $\mathbf{w}^t$ of the association potential in each epoch, the transition probabilities for the temporal association potential linking the map to the first epoch, and the transition probabilities between all pairs of neighboring epochs. For that purpose, the multitemporal CRF is split into M monotemporal CRFs. We start by training the classifier for epoch $t = 1$ using the iterative training and inference method previously described. Apart from the parameters of the CRF related to that epoch, this will also deliver the most likely class labels C_n^1 for the first epoch. These class labels can be used as observed class labels for monotemporal training and inference in epoch 2, which results in the CRF parameters related to epoch 2 and the class labels C_n^2, and so on. This procedure is repeated, for each epoch $t > 1$ treating the estimated class labels C_n^{t-1} as observed labels, until all parameters and all weights have been determined.

In the training process of the label noise robust logistic regression, a transition probability $p(\widetilde{C}_n \mid C_n)$ and the parameters of the posterior $p(C_n \mid \mathbf{x}_n)$ are estimated. The latter, based on data $\mathbf{x}^t$, is used for the association potential $A_x(C_n^t, \mathbf{x}^t) = \ln(p(C_n^t \mid \mathbf{x}_n^t))$.

For $t = 1$, the transition probability $p(\widetilde{C}_n \mid C_n^1)$ can be used directly to determine the temporal association potential for the observed map according to Equation 13. Similarly, the transition probabilities for the other epochs can be interpreted as the probabilities of inverse temporal change, $p(C_n^{t-1} \mid C_n^t)$, which, also assuming $p(C_n^t = C^k) \approx p(C_n^{t-1} = C^k)$, can be used to determine $p(C_n^t \mid C_n^{t-1})$ analogously to Equation 11. From the sequential training and classification the weights θ_n^t are defined for every epoch as well. They describe the areas of potential change between t and t–1. For $t = 1$, they are used as weights for the temporal association potential A_T in Equation 10, with $\widetilde{C}_n^0 = \widetilde{C}_n$ and $\theta_n^1 = \theta_n$. For all other epochs, they can also be used as weights $\vartheta_{n,i}^{t,t-1} = \vartheta_{i,n}^{t-1,t}$ for the temporal interaction potential, which, thus, becomes $I_T(C_n^t, C_i^{t-1}) = \vartheta_{n,i}^{t,t-1} \cdot \ln(p(C_n^t \mid C_i^{t-1}))$.

Having applied the training procedure just described, the class labels of all pixels in all epochs are already determined, for each epoch t only based on information from previous epochs. This can already be interpreted as the result of multitemporal processing, and it will be referred to as the *sequential multitemporal* CRF solution. Similarly to Hoberg *et al.* (2015), we also investigate a setting in which the results of the sequential approach are refined by a series of iterations of LBP where information is also passed on from epochs t to $t - 1$, using $I_T(C_i^{t-1}, C_n^t) = \vartheta_{i,n}^{t,t-1} \cdot \ln(p(C_i^{t-1} \mid C_n^t))$. This variant is referred to as the *fully multitemporal* CRF.

Experiments

Test Data and Test Setup

We used four datasets in our experiments. The first one consists of a part of the Vaihingen data of the ISPRS 2D semantic labeling contest (Wegner *et al.*, 2015). We use ten of the training patches, each consisting of about 2,000 × 2,500 pixels. For each patch, a color infrared true orthophoto (TOP) and a DSM are available with a ground sampling distance (GSD) of 9 cm. The reference consists of five classes: *impervious surfaces* (*sur.*), *building* (*build.*), *low vegetation* (*veg.*), *tree*, and *car*. As cars are not a part of a topographic map, this class was merged with *sur*. For each pixel, we defined a feature vector $\mathbf{x}_n(\mathbf{x})$ consisting of the normalized difference vegetation index (NDVI), the normalised DSM (nDSM), the red band of the TOP smoothed by a Gaussian filter with $\sigma = 2$, and hue and saturation obtained from the TOP, both smoothed by a Gaussian filter with $\sigma = 10$. These features were selected from a larger pool based on the feature importance analysis of a random forest classifier (Breiman, 2001).

The second and third datasets are based on satellite imagery (Maas *et al.*, 2016). The first one consists of a Landsat image from 2010 of an area near Herne, Germany, with a GSD of 30 m and a size of 362 × 330 pixels. The second dataset consists of a RapidEye image of an area near Husum, Germany, from 2010. Its GSD is 5 m and its size is 3,547 × 1,998 pixels. In both cases, only the red, green, and near infrared bands are available. The reference contains four classes *residential area* (*res.*), *rural streets*, *forest* (*for.*) and *cropland* (*crop.*). As the class *rural streets* is underrepresented in both images, we merged it with *cropland*. In both datasets, 19 features were selected to capture spectral information, texture and local context. As the resolutions of the Herne and Husum datasets are different, the window sizes for the definition of some of the features differ. First, we used five spectral features: near infrared band, intensity, hue, saturation, and NDVI. In Herne, we used the original features, whereas for Husum, they were smoothed by a Gaussian filter with $\sigma = 5$. We added the mean and variance of these features in a local neighborhood of 6 × 6 pixels for Husum and of 3 × 3 pixels for Herne. The set of features was completed by four Haralick features (energy, contrast, homogeneity, and entropy) related to texture, using a window of 5 × 5 pixels (Husum) and 3 × 3 pixels (Herne) to calculate the gray level co-occurrence matrices.

For the experiments with Vaihingen, Husum, and Herne we manually changed the reference to simulate an outdated map. For that purpose, we mainly deleted buildings from the reference and replaced them by a mixture of grass, trees, and roads that would be typical for open landscape to simulate typical patterns of urban development. However, we also changed some of the other classes to simulate other types of changes inside urban areas, e.g., by adding and deleting trees. In all cases, the simulated maps looked realistic. Thus, the outlines of the objects are changed in most cases, because, for instance, a building has another shape than a tree. For each patch of the Vaihingen dataset, three simulated maps were created, each with a different amount of change from about 5 percent up to about 35 percent. For Herne and Husum, the changed maps from Maas *et al.* (2016) were used. Of course, the original (unchanged) reference is used for evaluation.

The third dataset is based on three georeferenced Landsat images, showing Las Vegas, Nevada in the years 1991, 2000, and 2016. The images from 1991 and 2000 have a size of 2,626 × 2,249 pixels and a GSD of 28.495 m. The 2016 image has a GSD of 30 m and a size of 2,494 × 2,136 pixels. The outdated map is created by manually labeling three classes: *ground* (*gr.*), *residential area* (*res.*), and *water* (*wat.*) in a Landsat image from 1986 covering the same area and having a GSD of 60 m. The map was rescaled to have the same GSD and size as the image of 1991. The reference for each image was also generated by manual labeling, but these data are only used for evaluation. Figure 4 shows a comparison of land cover between the outdated map and the epoch 2016.

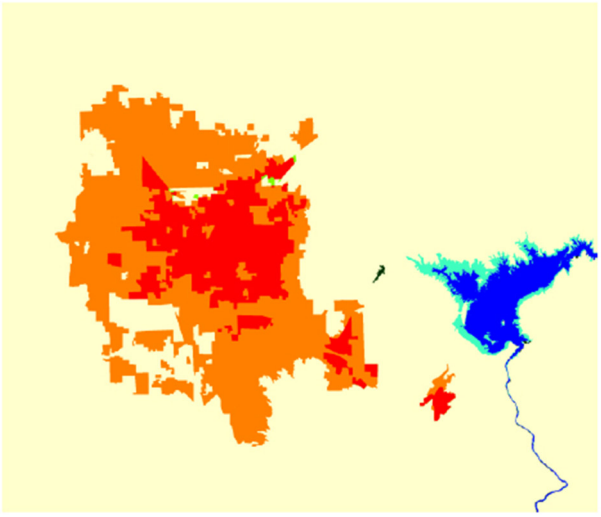


Figure 4. Comparison of the land cover in Las Vegas between 1986 and 2016. Unchanged areas - Red: *res*, yellow: *gr.*; blue: *wa*. Areas of change –orange: change from *gr*. to *res*.; cyan: change from *wa*. to *gr*.; dark green: change from *gr*. to *wa*.; light green: change from *res*. to *gr*. The most dominant type of change is from *gr*. to *res*. (orange), indicating the fast urbanization of that area.

For feature selection, the outdated map and the satellite data from 1991 were used. Nine pixel-based spectral features were used (blue, green, red, near infrared, and infrared band; NDVI; intensity, hue, and saturation, all from an RGB image). In addition, we determined 11 spectral features extracted from a local neighborhood (minimum, maximum, mean and variance of NDVI as well as intensity and saturation, all defined in a local window of 9 × 9 pixels; NDVI, intensity and saturation smoothed by a Gaussian filter with $\sigma = 4$). For textural information the four Haralick features energy, contrast, homogeneity and entropy were extracted based on a local window of 4 × 4 pixels. These 24 features defined a pool of features from which we selected the most relevant ones by a heuristic feature selection procedure. First, we selected two features (infrared band, NDVI) which resulted in a good separation of clusters for each class in feature space based on visual inspection. Using these two features, we classified the data from 1991 using label noise robust logistic regression without CRF and computed the kappa coefficient by comparing the classification results to the old map. This classification and evaluation procedure was repeated 22 times, each time using three features: the infrared band, the NDVI and one of the remaining features. The kappa coefficient is used to rank the remaining 22 features. For each feature type (pixel based and local spectral features, textural features), we selected the first two according to the ranking. The remaining feature vector, thus, only contains eight features: infrared band and NDVI, blue band and intensity; maximum and variance of NDVI in a local window of 9 × 9 pixels; Haralick contrast and homogeneity, both based on a window of 4 × 4 pixels.

For all experiments we used a feature space mapping $\mathbf{\Phi}(\mathbf{x}_n)$ based on quadratic expansion. The hyperparameter for regularization in Equation 3 was set to $\sigma = 10$. The initial values for the transition matrix $\mathbf{\Gamma}$ were $\gamma_{ij} = 0.8$ for $i = j$ and $\gamma_{ij} = 0.2/(K–1)$ for $i \neq j$, where K is the number of classes. The initial values for the parameter vector $\mathbf{w}$ of logistic regression were determined by standard logistic regression training without assuming label noise. The parameters of the contrast sensitive Potts model were set to β_0=1.0 and β_0=0.5.

Table 1 shows which heuristics were used for defining compact clusters of change according to the Training and Inference Section. The first two heuristics were found to be relevant for the Las Vegas data. For that dataset, applying heuristic 3 (related to small clusters of change) would have removed too many real changes, because areas of change were rather small; on the other hand, typical classification errors corresponded to larger clusters of *res.* and *wa.* pixels that in reality corresponded to *gr.* Furthermore, small gaps inside contiguous *res.* areas were frequently found to correspond to classification errors as well. Thus, the first two (class-specific) heuristics were found to be appropriate here, in case of heuristic 1 changing the class labels for obtaining the map B_C of potential change to *gr* for both, small *res.* and *wa.* clusters. For all the other datasets, heuristic 3 was found to work well, so the first two heuristics need not be applied. Heuristics 4 and 5 only make sense for very high-resolution data, thus they were only applied for processing the Vaihingen dataset.

The value c for updating the weights was found empirically and set to 0.1. Except for the dataset of Herne, where all pixels are used due to the small image size, not all available pixels are used for training to reduce the processing time. For Vaihingen just about 20 percent and for Las Vegas and Husum about 1 percent of the data remain in the training data.

The iteration is terminated if either less than 0.01 percent of the weights for the observed labels in classification change or if at least 40 iterations have been performed. In most cases, the iteration process converges so that the first criterion is responsible for the termination of the process. Even in cases when the second criterion became relevant, we found the number of pixels with changing weights to be so small that the subsequent analysis was hardly affected.

For all datasets we carried out four experiments. In the first experiment (**Init**), training and classification was carried out without iterative re-training and classification ($g_n = 1 = \text{const}$, $\theta_n = 0 = \text{const}$). This corresponds to a CRF that does not consider the observed labels from the map in classification, trained using the same weights $g_n = 1$ for all samples. The second experiment ($\mathbf{V}^g$) is based on our method, but without

Table 1. Parameters for the heuristics for determining compact clusters of change; the numbering scheme corresponds to one in the methodological section, where the heuristics are introduced.. The symbol - indicates that a heuristic was not applied, otherwise the parameter value is given, or the heuristic is marked by a "+".

	Heuristic		1 small objects	2 Small objects surrounded by another class	3 small clusters of change	4 object borders	5 shadow
	Parameter		o_k	z_k	u	s	
Dataset	Vaihingen		-	-	4×4m²	0.5m	+
	Herne		-	-	500×500m²	-	-
	Husum		-	-	500×500m²	-	-
	Las Vegas	*gr.*	-	-	-	-	-
		res.	1.5×1.5km²	250m			
		wa.	500×500m²	-			

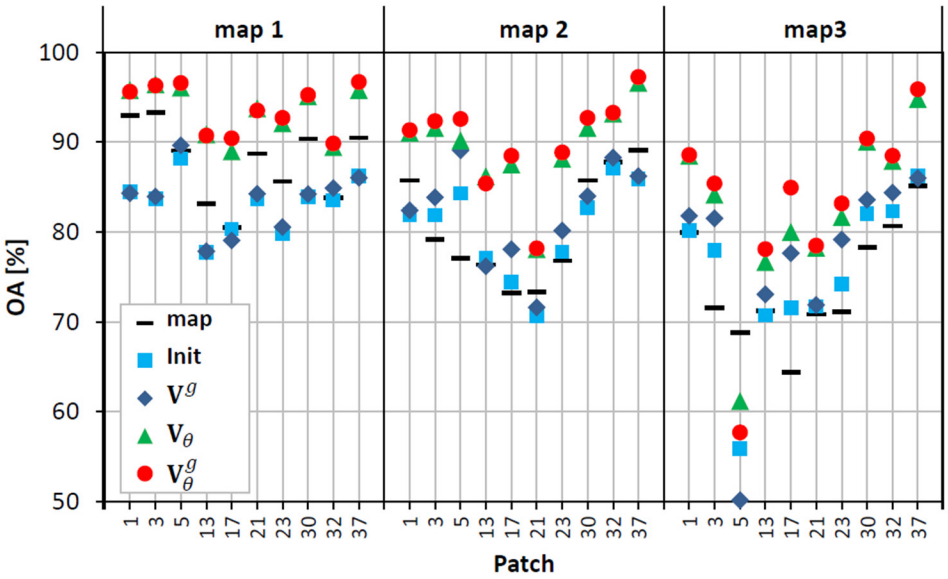


Figure 5. Overall accuracy for the four variants in Vaihingen and the outdated maps.

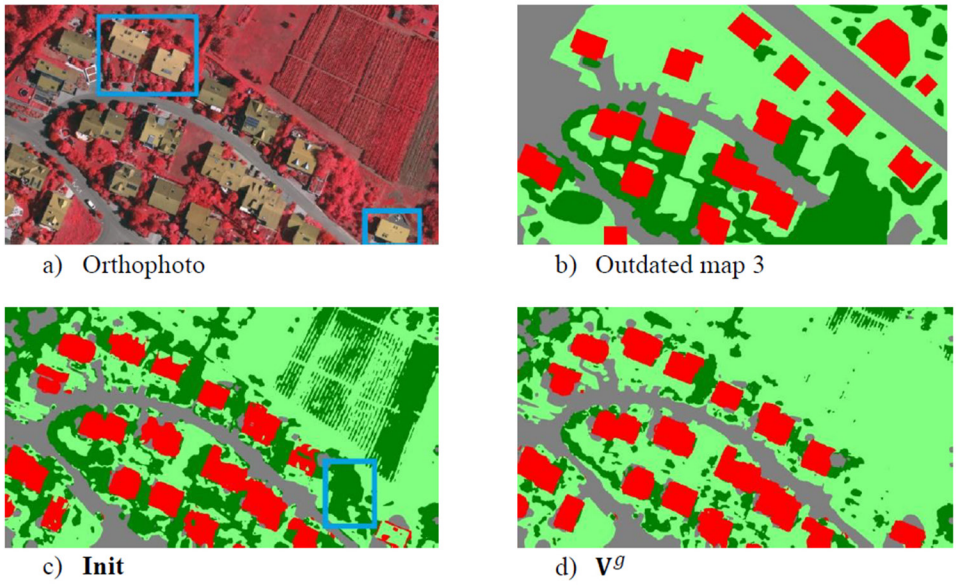


Figure 6. Data and results from patch 17 (Vaihingen). Red: *build.*, dark green: *tree*, light green: *veg.*, gray: *sur.*. Blue rectangles: highlighted objects discussed in the text.

considering the outdated map ($\theta_n = 0 = \text{const}$). It shows the impact of the sample weights g_n introduced in the Label Noise Robust Logistic Regression Section in the training step. The third experiment ($\mathbf{V}_\theta$) uses constant training weights $g_n = 1$, but does apply the modified weights θ_n to include the map information. The last experiment ($\mathbf{V}_\theta^g$) uses our method with weight modification both in the training and classification steps. In each case, we compare the results to the reference on a per-pixel basis, determining the overall accuracy (OA) as well as completeness and correctness per class (Heipke *et al.*, 1997). Comparing the simulated map to the real reference shows the amount of change in the corresponding dataset, and the resultant quality indices are also reported (**map**); 100% - OA of **map** gives the amount of simulated change in each experiment. We do not distinguish a training set from a test set because an outdated map is always used, at least for training. However, the reference used for the evaluation is different from the outdated map in areas with simulated changes.

For the Las Vegas dataset we can distinguish between three different settings. Firstly, we can apply both the sequential multitemporal CRF ($^S\mathbf{V}$) and the fully multitemporal CRF ($^M\mathbf{V}$). Furthermore, we can apply monotemporal CRF ($^N\mathbf{V}$), c.f. Section 3.2, just using the old map from 1986 and the most current data from 2016, assuming no image data to be given for the intermediate epochs. In all three cases ($^M\mathbf{V}$, $^S\mathbf{V}$ and $^N\mathbf{V}$), experiments are conducted based on the four variants **Init**, $\mathbf{V}^g$, $\mathbf{V}_\theta$, and $\mathbf{V}_\theta^g$ introduced previously. In these experiments, only the map from 1986 is used in the training process, the other maps are only used for evaluation.

Results and Evaluation

Vaihingen

Figure 5 shows the OA of all patches achieved for three versions of the outdated map for Vaihingen. The average amount of label noise was 12%, 20%, 26% for maps 1, 2 and 3, respectively. In most cases, the variant $\mathbf{V}_\theta^g$ achieves the best OA (85%-90%), but variant $\mathbf{V}_\theta$ performs at a similar level, and both variants clearly outperform the variants without weights and without considering the outdated map (**Init**, $\mathbf{V}^g$). Obviously, the inclusion of the outdated map has a relatively high impact on the quality of the results, improving the OA by 2%-10%. The mean OA increases by 10% for map 1, 9% for map 2 and 7% for map 3. This is mainly caused by an improved classification at object boundaries or at individual pixels. In fact, in some cases, the variants not considering the outdated map in classification (**Init**, $\mathbf{V}^g$) lead to results where a larger percentage of change than actually present is predicted, so that the corresponding OA is lower than the one indicated by **map**.

The advantage of considering the sample weights g_n in training becomes more obvious for experiments with a large amount of change. If the level of change is small (map 1), it can be compensated by the original method based on the NAR model (**Init**; Maas *et al.*, 2016). If the label noise cannot be compensated by the NAR model anymore, considering the weights can improve the results. One example is patch 17 (Figure 6). It contains three buildings with a brighter appearance than the rest (blue rectangles in Figure 6a). Only one of them is contained in the outdated map (Figure 6b). Without considering the weights (variant **Init**, Figure 6c), one building is mostly classified as *veg.* In variant $\mathbf{V}^g$, the two changed buildings are correctly detected (fig. 6(d)). Another difference

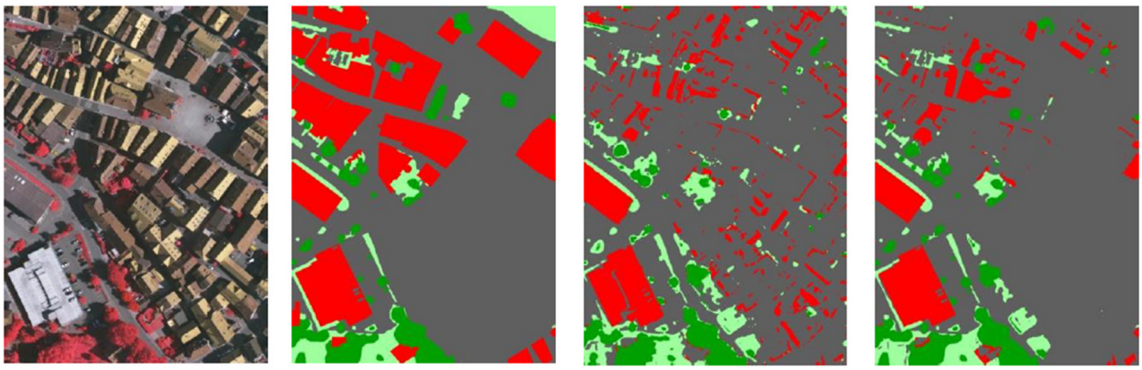


Figure 7. Data and results from patch 5 (Vaihingen). Red: *build.*, dark green: *tree*, light green: *veg.*, gray: *sur.*

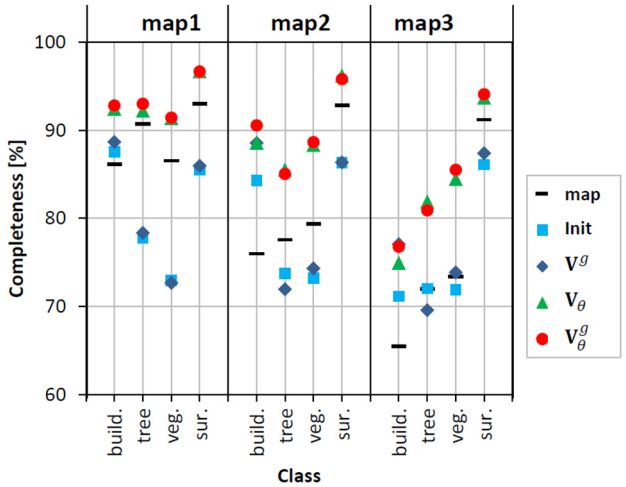


Figure 8. Average completeness over all patches in Vaihingen.

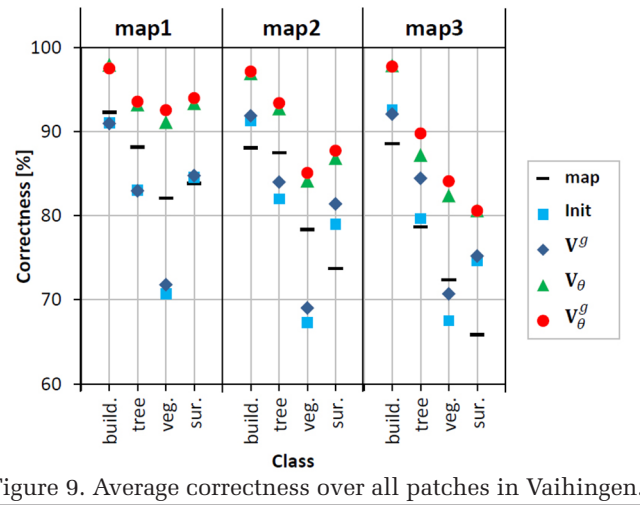


Figure 9. Average correctness over all patches in Vaihingen.

between the results of variants **Init** and $\mathbf{V}^g$ is the label of the vineyard which belongs to the class *veg.* but is often classified as *tree* in experiment **Init**.

Without considering weights, the probability $p(C_n \mid \mathbf{x}_n)$ is low for all classes in the area of the vineyard, so that the classification results are not reliable. By considering the sample weights in experiment $\mathbf{V}^g$, the probability $p(C_n \mid \mathbf{x}_n)$ for the class *veg.* is much higher than for the other classes. However, because the vineyard has a similar appearance to trees, the tree marked by a blue rectangle in Figure 6c is also classified as *veg.* in $\mathbf{V}^g$.

For patch 5 (Figure 7) and the outdated map 3 with more than 30 percent label noise, OA is always below 61 percent (Figure 5). In this case, nearly 50 percent of all building pixels are labeled as *sur.* in the outdated map. This amount of label noise cannot be dealt with by the original method (**Init**). The transition probabilities γ_{ii} for no change for *build.* and *sur.* determined in the initial training step are close to 1 and, thus, not very accurate. Consequently, the iterative weight updating procedure does not converge to the correct solution. In summary, a limitation of the algorithm is that for each cluster in feature space enough correctly labeled samples must be part of the training dataset created from the outdated map.

Figures 8 and 9 show the completeness and the correctness of the results. Both quality indices are higher for variants $\mathbf{V}_\theta$ and $\mathbf{V}^g$ than for the others, which again highlights the importance of using the outdated map for classification. Using the sample weights g_n in the training process does not improve the completeness in most cases, but it does have a small positive impact on the correctness.

For buildings, we also provide an evaluation on a per-object basis, counting a detected building (i.e., a connected component of pixels classified as *build.*) as a true positive if more than 70 percent of its area overlaps with a reference building. Because small buildings are often not included in maps, buildings smaller than 16 m² were excluded from the evaluation. The mean completeness and correctness of all areas are shown in Table 2. Again, variant $\mathbf{V}_\theta^g$ achieves the best completeness (98.9 percent) and correctness (82.5 percent). However, variants $\mathbf{V}_\theta$ and $\mathbf{V}^g$ do not perform significantly worse considering the standard deviations of the quality indices. Nevertheless one can notice a positive impact of the new developments presented in this paper (variants $\mathbf{V}^g$, $\mathbf{V}_\theta$ and, particularly, $\mathbf{V}_\theta^g$) compared to the original algorithm (**Init**) (Maas et al., 2016).

Table 2. Completeness and correctness on a per-object basis for buildings (Vaihingen); mean of all areas in % [standard deviation in %].

	$\mathbf{V}_\theta^g$	$\mathbf{V}_\theta$	$\mathbf{V}^g$	**Init**	**map**
Corr.	99 [4]	99 [4]	96 [8]	93 [14]	91 [11]
Comp.	82 [19]	80 [17]	82 [18]	75 [20]	74 [13]

Husum and Herme

As the amount of change in Husum and Herne is quite small (3 percent - 4 percent), using the sample weights g_n does not affect the results much; the OA changes by less than 0.6 percent. Thus, this section focuses on the impact of using the

outdated map for classification. In Table 3 OA, completeness and correctness are shown for both datasets for variants **Init** and $\mathbf{V}_\theta$. All values are larger for variant $\mathbf{V}_\theta$ by a large margin, the OA increasing by 11.4 percent for Herne and by 5.1 percent for Husum. One reason for that increase is the improvement of the delineation of object borders and the detection of small objects. As the features depend on a local subset of pixels, borders of objects are blurred in the standard classification process. In variant $\mathbf{V}_\theta$, these areas can be correctly classified in regions without change. To highlight the potential for detecting changes despite using the existing map for classification, Table 4 shows the OA achieved for pixels in the areas affected by a change. The results show that the improved OA for the entire image caused by the inclusion of the outdated map (cf. Table 3) comes at the cost of a reduced OA in the changed areas. In Husum, this reduction in OA is low (0.4 percent). In Herne it is larger (3.4 percent), though still considerably smaller than the improvement for the entire scene (11.4 percent). Nevertheless, it is obvious that the vast majority of changed pixels is classified correctly even in the setting using the existing map. As the OA is increased considerably, it is also obvious that this comes along with a considerable reduction in false alarms in unchanged areas.

Table 3. OA, completeness, correctness for Husum and Herne [%].

		Herne		Husum	
		$\mathbf{V}_\theta$	**Init**	$\mathbf{V}_\theta$	**Init**
OA		94.8	83.4	98.5	93.4
Compl.	res.	96.3	88.0	91.3	79.1
	for.	91.5	69.5	95.7	77.8
	crop.	95.5	87.7	99.7	96.8
Corr.	res.	92.5	81.8	96.7	81.1
	for.	97.5	84.1	98.9	84.1
	crop.	95.2	84.4	98.6	95.8

Table 4. OA of Husum and Herne for areas affected by a change [%].

Herne		Husum	
Init	$\mathbf{V}_\theta$	**Init**	$\mathbf{V}_\theta$
85.4	82.0	95.4	95.0

Las Vegas

Figure 10 shows the OA for all experiments and all epochs of the Las Vegas data set. The variant that is comparable to the results achieved for the other datasets, based only on the outdated map from 1986 and the image from 2016, is variant ${}^{N}\mathbf{V}$ (indicated by the blue bars in Figure 10). In this setting, the highest OA is achieved by using variant $\mathbf{V}_\theta$ with 95.6 percent, closely followed by variant $\mathbf{V}_\theta^g$, resulting in an OA of 95.5 percent, whereas the original method of Maas *et al.* (2016) (**Init**) results in the smallest OA (94.9 percent). Variant $\mathbf{V}^g$ achieves an OA of 95.0 percent, which is slightly lower than the one achieved for $\mathbf{V}_\theta^g$. This behavior is similar to Vaihingen, Husum, and Herne. Analyzing the transition probabilities $p(\tilde{C} \mid C)$ in the training procedure of label noise robust logistic regression shows that the large transition from *residential* to *ground* in the map $p(\tilde{C}=C^{gr} \mid C=C^{res})$ is estimated quite well (69 percent versus 72 percent in the reference). Small transition probabilities of ground and water are not captured in the training. Despite the large time difference of 30 years and a considerable amount of change caused by the urban sprawl of Las Vegas in that period, the results achieved in this setting that is only based on one image without any additional training data beyond the existing land cover map are very encouraging.

In order to analyze the multitemporal case, we also consider the epochs of 1991 and 2000; the OA are also displayed in Figure 10, using red color for the sequential multitemporal (${}^{S}\mathbf{V}$) and green for the fully multitemporal CRF (${}^{M}\mathbf{V}$). Note that for the fully multitemporal versions, ${}^{M}\mathbf{V}$, in variants $\mathbf{V}^g$ and **Init**, we used temporal transitions (otherwise the concept of a multitemporal CRF would not make sense), but did not use individual weights for the temporal interaction potentials, setting $\vartheta_{i,n}^{t,t-1} = 1$ for all pixels and epochs.

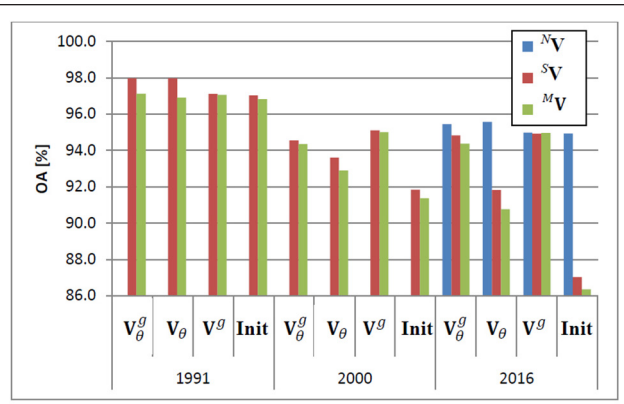


Figure 10. Overall accuracy (OA) for all variants in all epochs of the Las Vegas dataset. Note that variant NV is based on the data from 2016 only, so that there are no blue bars for the other epochs.

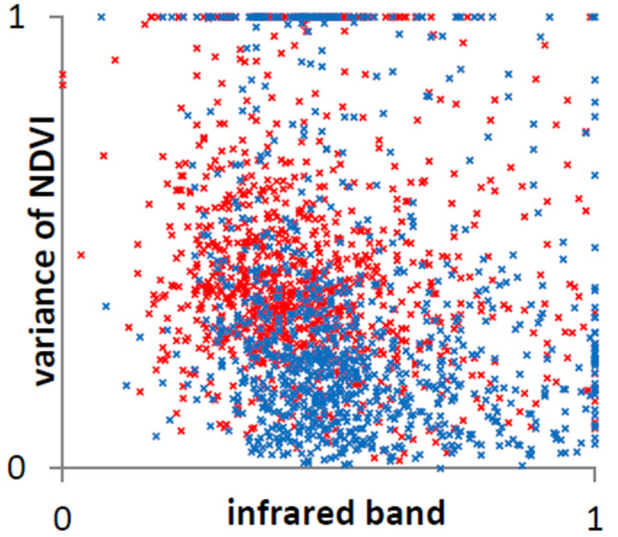


Figure 11. Scatterplot of infrared band and the variance of NDVI for the epoch 2000. The plot only shows 5 percent of the pixels for better readability. Red: unchanged res. pixels (i.e., pixels corresponding to res. both in 1991 and in 2001); blue: pixels changed from gr. (1991) to res. (2000). The figure indicates that the distributions of the features for settlement areas developed in that period (blue) is slightly different from the distribution of the features for the unchanged pixels.

A first glance at Figure 10 shows that in general, the classification results are good independent from the method that was used. In all variants except **Init** for the multitemporal settings in 2016, the OA is higher than 90 percent. In general, the OA is higher for epoch 1991 than for the other ones. This is not surprising, because the time difference between the original map (1986) and that epoch is relatively low, so that the amount of change is not very large: only 2 percent of the pixels changed their class label between these epochs. Note that for the first year, variant ${}^{S}\mathbf{V}$ corresponds to the setting where only one image and the original map are available, and similarly to the results for Vaihingen, Herne, and Husum, the improvement in OA caused by using the existing maps ($\mathbf{V}_\theta$, $\mathbf{V}_\theta^g$) is larger than the improvement due to reducing the impact of potentially wrong training data ($\mathbf{V}^g$). The estimation of the transition probabilities was quite accurate, which explains the positive impact of these weights on the results.

For 2000, the OA is considerably lower than for 1991 in all variants. In this period, the amount of change was much larger than between 1986 and 1991 (7 percent of the pixels changed their class labels), mainly due to an expansion of the residential area: according to the reference, 47 percent of the *res.* pixels in 2000 had still been *gr.* in 1991. In addition, the appearance of the new buildings in the imagery was slightly different, which is illustrated by the scatterplot in Figure 11. As a consequence, label noise tolerant training of the logistic regression classifier does not work as well as for the previous period, in particular also giving a relative poor estimate for the transition probabilities. Especially the transition from *residential* to *ground* in the map $p(\tilde{C}=C^{gr} \mid C=C^{res})$ is estimated too low (e.g. 8.7 percent versus 46 percent in the reference for variant **Init**). In such a setting, including the (not very precise) transition probabilities may have a smaller positive effect than reducing the impact of (a considerable number) of wrong training samples. This leads to a situation where the weighting of potentially wrong training samples ($\mathbf{V}^g$, $\mathbf{V}^g_\theta$) has a higher positive impact on the results than using individual weights for the transition probabilities in the temporal model ($\mathbf{V}_\theta$).

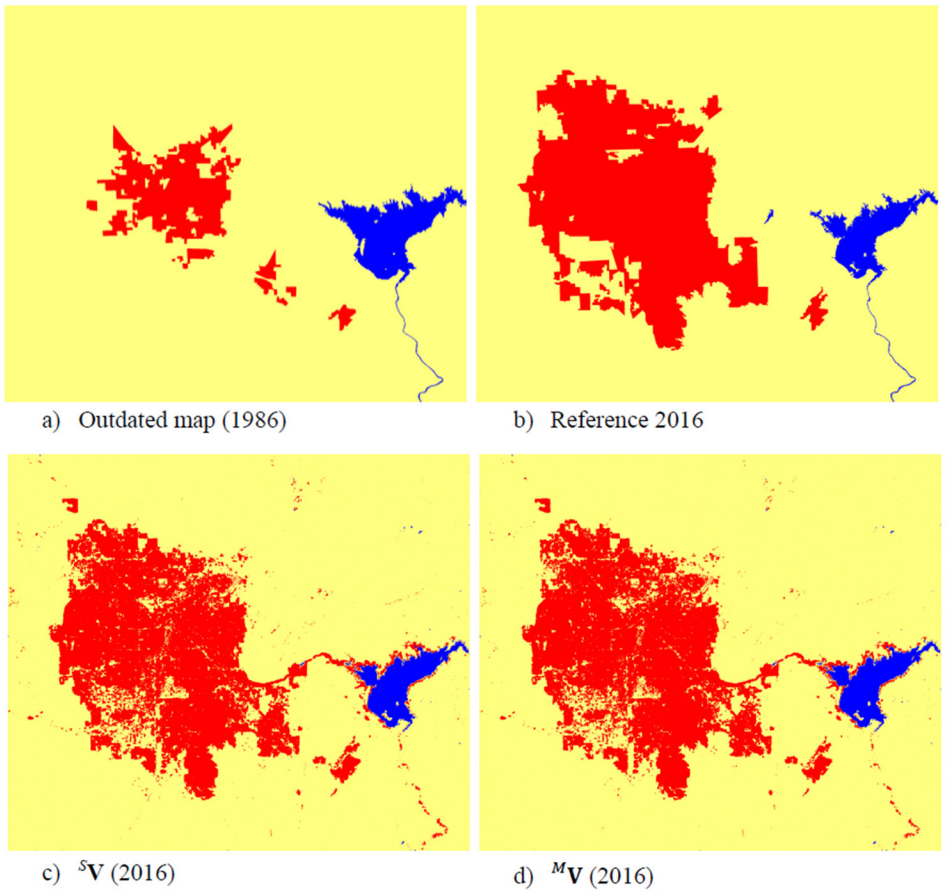


Figure 12. Outdated map (a), reference of epoch 2016 (b), and the resulting label images of variant $\mathbf{V}^g$ for 2016, using the sequential ($^S\mathbf{V}$, c) and the fully multitemporal CRF ($^M\mathbf{V}$, d). Colors: red – *res.*, yellow – *gr.*, blue – *wa*.

Finally, for epochs 2016, the OA remains on a similar level as for 2000, despite the fact that the overall amount of change between these periods is also about 7%, and 31% of the *res.* pixels from 2016 were still *gr.* in 2000. Here, the transition probabilities were estimated in a better way, but training for the potentials related to 2016 being based on the classification results of 2000, the method cannot recover from errors committed earlier, and it would seem that the variations between the variants in 2016 largely reflect the variations also occurring in 2000.

Comparing the different variants of training, it becomes obvious that the original method (**Init**, Maas *et al.*, 2016) performs considerably worse than the other ones, and in a setting with large changes and changing appearance of the data, variants based on weights for the training samples ($\mathbf{V}^g$, $\mathbf{V}^g_\theta$) are preferable to variant $\mathbf{V}_\theta$. A comparison between $\mathbf{V}^g$ and $\mathbf{V}^g_\theta$ does not lead to such obvious conclusions: $\mathbf{V}^g_\theta$ performs better than $\mathbf{V}^g$ in epoch 1991 (where the transition probabilities are estimated quite well) for the sequential multitemporal CRF by a margin of about 1%, but in the other epochs $\mathbf{V}^g$ outperforms $\mathbf{V}^g_\theta$, though only by a small margin.

The differences between the two multitemporal settings are very small in most cases, the exception again being epoch 1991, where the sequential setting ($^S\mathbf{V}$) outperforms the fully multitemporal one ($^M\mathbf{V}$) by a margin of 1 percent. The main difference between these settings is that in $^M\mathbf{V}$, another inference process is performed on top of the processing chain corresponding to $^S\mathbf{V}$ in which temporal information is passed on in two directions. The results in Figure 10 indicate that this additional information flow does not improve the OA, and in 1991, it obviously leads to a slight degradation of performance, because the errors in the transition probabilities from 1991 to 2000 also propagate wrong information from 2000 to 1991. Figure 12 shows the label images for 2016 achieved using the two settings $^S\mathbf{V}$ and $^M\mathbf{V}$, in both cases based on variant $\mathbf{V}^g$, in which individual weights are applied for training samples, but not for the temporal interactions. The figure also shows the outdated map and the reference for 2016.

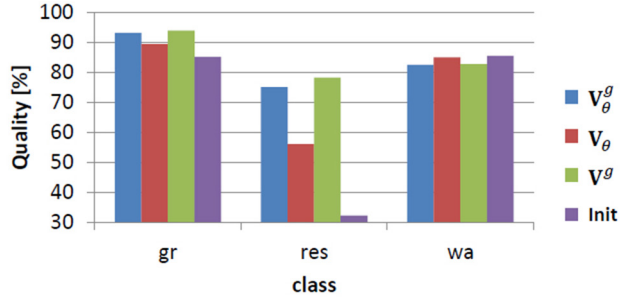


Figure 13. Quality of each class for the epoch 2016 in the fully multitemporal setting ($^M\mathbf{V}$).

The observations made for overall accuracy are also reflected by a closer inspection of class-wise accuracy indices. As an example, Figure 13 shows the quality, a parameter considering both false positive and false negative pixels of that class (Heipke *et al.*, 1997), for each class in epoch 2016, achieved

using the fully multitemporal setting $^{M}\mathbf{V}$; the corresponding numbers for the other settings and epochs follow a similar pattern and are omitted for the sake of brevity. In general, quality is quite high, and one can see that the different variants mainly differ for class *res.*, the class affected by the largest overall amount of change in the sense that a large majority of the *res.* pixels in 2016 had not belonged to that class in 1986. Figure 13 indicates that the differences between the variants discussed above mainly affect class *res.*, and it becomes obvious that the quality for that specific class can be improved dramatically if individual weights for training samples are considered: variant $\mathbf{V}^{g}$ leads to an improvement of more than 20 percent in quality for class *res.* compared to $\mathbf{V}_{\theta}$.

The label images obtained for epoch 2016 using the fully multitemporal setting ($^{M}\mathbf{V}$) for variants **Init**, $\mathbf{V}_{\theta}$ and $\mathbf{V}_{\theta}^{g}$ are shown in Figure 14; variant is shown in Figure 12d. These figures also indicate that considering individual weights in training (cf. Figure 12d and Figure 14c as well as the reference in Figure 12a) increases the accuracy of detection of residential areas, especially newly built-up areas at the border of the city.

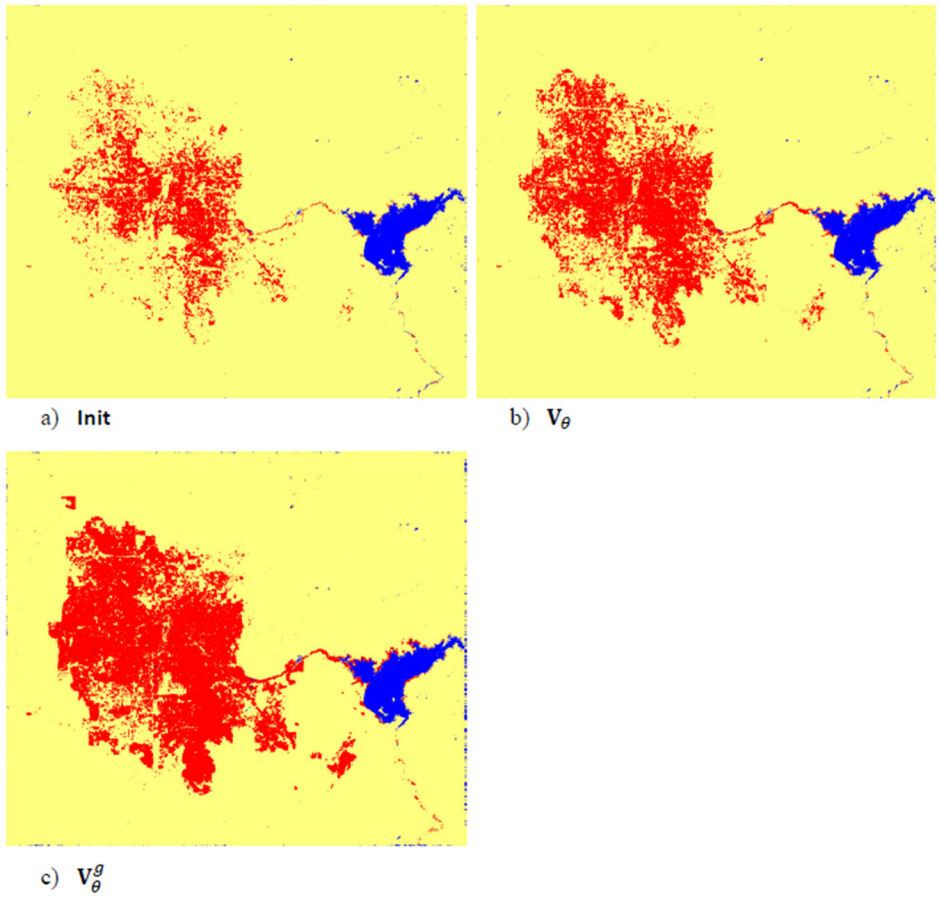


Figure 14. Label images of three variants using the fully multitemporal CRF ($^{M}\mathbf{V}$); the outdated map, the reference, and variant $\mathbf{V}^{g}$ are shown in Figure 12a, 12b, and 12c, respectively. The color code is identical to the one used in Figure 12.

This impression is further supported by Table 5, which presents OA for 2016 just considering the pixels affected by a change (i.e., only considering pixels whose class label in the outdated map from 1986 is different from the one in the reference of 2016. This would be about 15 percent of all pixels, most of which would correspond to class *res.* in 2016). The table shows that in variant **Init**, only 27 percent of all changes can be detected in the fully multitemporal setting. Considering individual weights for the temporal transitions ($\mathbf{V}_{\theta}$) helps, but only when training weights are considered ($\mathbf{V}_{\theta}^{g}$), more than 77 percent of the changes can be detected correctly. In case of variant $\mathbf{V}^{g}$, the OA for changes is even larger: 83 percent of the changes can be detected correctly.

Figure 15 shows the corresponding completeness and correctness values for classes *gr.* and *res.* in the areas affected by a change between 1986 and 2016 (cf. fig. 5); *wa.* is not analyzed because it is underrepresented in the areas affected by change. The figure shows that the correctness for *res.* is close to 100% for all variants, its completeness varying between 21% (variant **Init**) and 81% (variant $\mathbf{V}^{g}$). For the completeness of *res.* in the changed areas, the consideration of training weights would have a very large positive impact, whereas considering the outdated map leads to a decrease of completeness of that class compared to the best method. The completeness of *gr.* is higher than 60% in all cases, with an advantage for variants **Init** and $\mathbf{V}_{\theta}$, but the correctness of that class is extremely low (the best variant, $\mathbf{V}^{g}$, only achieving 21%). It would seem that the classification of *gr.* areas is very uncertain. As a large majority of change is from *gr.* to *res.*, the OA in Table 5 is dominated by the quality of the results

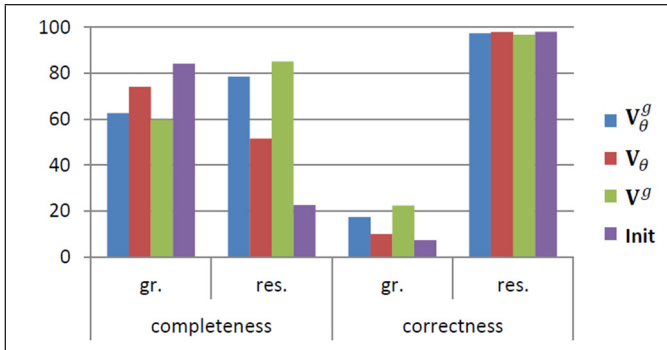


Figure 15. Completeness and correctness of the classes *gr.* and *res.* for the fully multitemporal CRF ($^{M}\mathbf{V}$) in epoch 2016 just considering pixels affected by any changed between epochs 1986 and 2016.

Table 5. OA achieved for the fully multitemporal CRF ($^{M}\mathbf{V}$) in epoch 2016 just considering pixels affected by any changed between epochs 1986 and 2016.

$\mathbf{V}_{\theta}^{g}$	$\mathbf{V}_{\theta}$	$\mathbf{V}^{g}$	**Init**
77.3%	53.1%	83.3%	26.8%

for *res.* (because this type of change would lead to class *res.* in the reference of the changed areas). There are only very few pixels which change from any class to *gr.*, so that the poor quality indices for that class have almost no influence

on the OA in the changed areas. In particular, we suspect the low correctness of *gr.* in the changed areas to be caused by missed new *res.* areas that were erroneously classified as *gr.* We conclude that the most frequent type of change can be detected relatively well by our methodology, although the inclusion of the outdated map leads to a certain amount of what we could call temporal oversmoothing.

To gain further insights into the behavior of the methods proposed in this paper, we also carried out an evaluation of their potential to detect changes. In this evaluation, a pixel of the image at epoch *t* was considered to be changed if its class label was different from the class label of the original map (1986); it was counted as a true positive if it corresponded to such a change both in the reference and in the classification result. The resultant quality indices are shown in Figure 16 for the multitemporal setting; the sequential mono-temporal setting shows a similar behavior and is omitted for the sake of brevity. In all cases, the inclusion of weights outperforms the variant **Init** (Maas *et al.*, 2016). As we have seen earlier, the integration of weights for the training samples improves the detection of new residential areas and, consequently, has a very large positive effect on the detection of changes; both the completeness and the correctness of variants $\mathbf{V}^g$ and $\mathbf{V}^g_\theta$ are much larger than for variants $\mathbf{V}_\theta$ and **Init** (Figure 16). The integration of the outdated map ($\mathbf{V}_\theta$, $\mathbf{V}^g_\theta$) leads to a reduction of all quality indices, most notably in the completeness of detected changes, compared to variant $\mathbf{V}^g$. The reduction is relatively low when weights are integrated in the training process ($\mathbf{V}^g_\theta$), but quite large when this is not the case ($\mathbf{V}_\theta$). Again, this shows that the integration of the outdated map results in a certain degree of temporal oversmoothing (indicated by the lower completeness of changed areas).

Finally, we want to compare the results of the multitemporal settings (${}^S\mathbf{V}$, ${}^M\mathbf{V}$) to those of the setting in which only epoch 2016 and the outdated map are considered (${}^N\mathbf{V}$). This comparison is based on Figure 10 again. The figure shows that for the best variants ($\mathbf{V}^g$, $\mathbf{V}^g_\theta$), the OA of the three settings is nearly identical. Thus, the additional data for the intermediate epochs do not increase the classification accuracy for the last epoch in the current implementation of the method. We think that this is because up to now there is no joint estimation of the transition probabilities and the parameters of the association potentials in the individual epochs. It is obvious that the sequential training procedure must lead to a degradation of results by accumulating errors. On the other hand, at least this accumulation of errors does not lead to worse results than disregarding the data from intermediate epochs if individual weights for training samples are used; the results for variant **Init** for 2016 show that in this simple setting, errors can accumulate so that the additional data actually lead to a considerable deterioration of OA. In any case, the multitemporal settings give access to monitoring the temporal evolvement of change by providing land cover information at intermediate epochs.

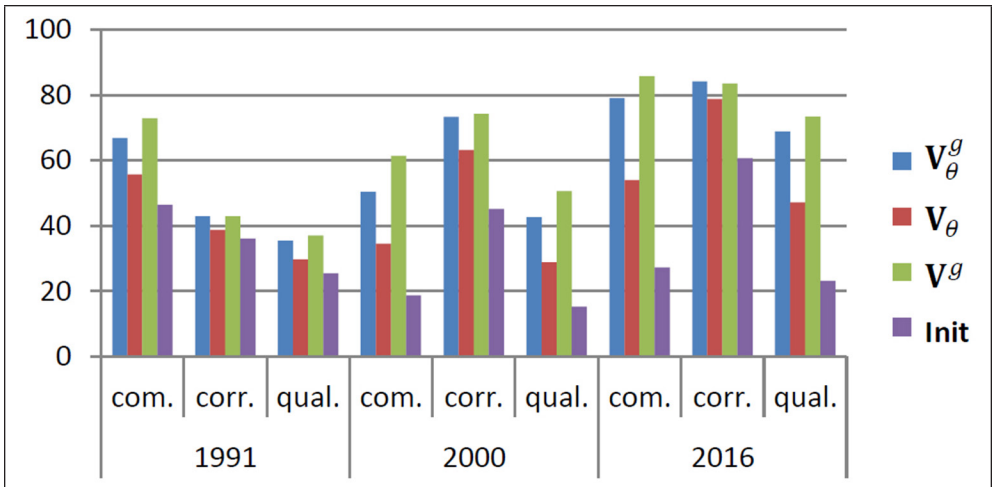


Figure 16. Completeness, correctness and quality of changes between 1986 and 2016 for the multitemporal setting.

Conclusions and Future Work

In this paper, we have presented an iterative method for supervised classification under label noise making use of the existing map both for training and in the classification process. No manual effort for the generation of training data was required. In both, the training and the classification procedure, we considered the fact that changes in land cover usually appear in clusters. In training, this was achieved by using a weight for each training sample in order to reduce the impact of samples in larger areas of change. By adding the labels of the map to the CRF as weighted observations, our method includes the map information for pixels that are unlikely to correspond to changes. Thus, new objects can be found without the additional map information while pixels probably not affected by label noise can take advantage of this prior information. Additional layers in the CRF model make it possible to consider data from more than one epoch and classify all epochs simultaneously in one multitemporal model. In this multitemporal setting, the parameters of the model for temporal transition can also be determined from the data without additional training labels.

We tested our method using datasets having different properties and varying degrees of label noise. Due to our re-weighting scheme for training samples the method can also deal with larger amount of noise; this innovation was particularly useful in the experiments using data of coarse resolution with a considerable amount of change that also involved changing appearance of classes in the data. The inclusion of the map information to the CRF has a considerably larger positive effect with high-resolution data and in scenarios where the changes do not go along with changes in appearance, largely due to a better classification of pixels near object boundaries. The actual changes are detected quite well, though small changed objects might not be detected and the evaluation for the multitemporal setting showed that the inclusion of the map into the classification process may lead to a certain degree of temporal over-smoothing. A major limitation of the method is that each cluster in feature space still must contain enough correct training samples for it to work. If the results of the base classifier in the initialization step are sufficiently good, considering the map in the classification can improve the results considerably. Using additional data from intermediate epochs in the multitemporal CRF did not improve the results for the classification of the last epoch compared to a variant without these data, while still giving

access to the temporal evolvement of change by providing label images for the intermediate epochs.

In our future work we want to expand our experiments to high resolution data with real changes to see how our method works under more realistic circumstances in terms of the extent of change, level of detail or number of classes for such data. We also want to see if the exchange of the base classifier, in this case the logistic regression, might have a positive effect on the results as well. In the multitemporal CRF, we would like to develop a joint training procedure that uses the results of the procedure presented in this paper as initial values. For instance, after initializing the CRF parameters, the class labels from the last epoch could be propagated back to the original map and differences between the original map and the propagated labels could be used to define a loss function for adapting the parameters of the temporal model.

Acknowledgments

The Vaihingen data set was provided by the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF) (Cramer, 2010):URL: *http://www.ifp.uni-stuttgart.de/dgpf/DKEP-Allg.html*. This work was supported by the German Science Foundation (DFG) under grant HE 1822/35-1.

References

An,W., and M. Liang, 2013. Fuzzy support vector machine based on within-class scatter for classification problems with outliers or noises, *Neurocomputing*, 110:101–110.

Bishop, C.M., 2006. *Pattern Recognition and Machine Learning*, First edition, Springer, New York.

Bootkrajang, J., and A. Kabán, 2012. Label-noise robust logistic regression and its applications, *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, pp. 143–158.

Boykov, Y., O. Veksler, and R. Zabih, 2001. Fast approximate energy minimization via graph cuts, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239.

Breiman, L., 2001. Random forests, *Machine learning*, 45(1):5–32.

Bruzzone, L., and C. Persello, 2009. A novel context-sensitive semisupervised svm classifier robust to mislabeled training samples, *IEEE Transactions on Geoscience and Remote Sensing* 47(7):2142–2154.

Bruzzone, L., D.F. Prieto, and S.B. Serpico, 1999. A neural-statistical approach to multitemporal and multisource remote-sensing image classification, *IEEE Transactions on Geoscience and Remote Sensing*, 37(3):1350–1359.

Büschenfeld, T., 2013. Klassifikation von Satellitenbildern unter Ausnutzung von Klassifikationsunsicherheiten, PhD thesis, Fortschritt-Berichte VDI, Reihe 10 Informatik / Kommunikation, Vol. 828, Institute of Information Processing, Leibniz Universität Hannover, Germany.

Cramer, M., 2010. The DGPF-test on digital airborne camera evaluation - Overview and test design, *Photogrammetrie-Fernerkundung-Geoinformation*, 2010(2):73–82.

Förstner, W., and B.P. Wrobel, 2016. Robust estimation and outlier detection, *Photogrammetric Computer Vision*, First edition, Springer, Cham, Switzerland, Chapter 4.7, pp. 141–159.

Frénay, B., and M. Verleysen, 2014. Classification in the presence of label noise: A survey, *IEEE Transactions on Neural Networks and Learning Systems* 25(5):845–869.

Frey, B.J., and D.J. MacKay, 1998. A revolution: Belief propagation in graphs with cycles, *Advances in Neural Information Processing Systems* (NIPS) 10: 479–485.

Heipke, C., H. Mayer, . Wiedemann, and O. Jamet, 1997. Evaluation of automatic road extraction, *International Archives of Photogrammetry and Remote Sensing*, Vol. XXII-3/4W2:151–160.

Hoberg, T., F. Rottensteiner, R.Q. Feitosa, and C. Heipke, 2015. Conditional random fields for multitemporal and multiscale classification of optical satellite imagery, *IEEE Transactions on Geoscience and Remote Sensing* 53(2):659–673.

Jia, K., S. Liang, X. Wei, L. Zhang, Y. Yao, and S. Gao, 2014. Automatic land-cover update approach integrating iterative training sample selection and a markov random field model, *Remote Sensing Letters*, 5(2):148–156.

Jiang-wen Sun, Feng-ying Zhao, Chong-jun Wang, and Shi-fu Chen, 2007. Identifying and correcting mislabeled training instances, *Proceedings of the Conference on Future Generation Communication and Networking* - Volume 1, pages 244–250.

Jianya, G., S. Haigang, M. Guorui, and Z. Qiming, 2008. A review of multi-temporal remote sensing data change detection algorithms, *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. XXXVII-B7, pp. 757–762.

Kumar, S., and M. Hebert, 2006. Discriminative random fields, *International Journal of Computer Vision* 68(2):179–201.

Li, Y., L.F. Wessels, D. de Ridder, and M.J. Reinders, 2007. Classification in the presence of class noise using a probabilistic kernel fisher method, *Pattern Recognition*, 40(12): 3349– 3357.

Lu, D. P. Mausel, E. Brondizio, and E. Moran, 2004. Change detection techniques, *International Journal of Remote Sensing*, 25 (12):2365–2401.

Maas, A., F. Rottensteiner, and C. Heipke, 2016. Using label noise robust logistic regression for automated updating of topographic geospatial databases, *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. III-7:133–140.

Maas, A., F. Rottensteiner, C. Heipke, 2017. Classification under label noise based on outdated maps. ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences, 4.

Mas, J.F., 1999. Monitoring land-cover changes: A comparison of change detection techniques, *International Journal of Remote Sensing*, 20 (1):139–152.

Melgani, F.; and S.B. Serpico, 2003. A Markov random field approach to spatiotemporal contextual image classification, *IEEE Transactions on Geoscience and Remote Sensing*, 41(11):2478–2487.

Mnih, V., and G.E. Hinton, 2012. Learning to label aerial images from noisy data, *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pp. 567–574.

Patrini, G., A. Rozza, A.K. Menon, R. Nock, L. Qu, 2017. Making deep neural networks robust to label noise: A loss correction approach, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR)*, pp 1944–1952.

Pelletier, C., S. Valero, J. Inglada, N. Champion, C. Marais Sicre, and G. Dedieu, 2017. Effect of training class label noise on classification performances for land cover mapping with satellite image time series, *Remote Sensing*, 9(2):173 –197

Radoux, J., and P. Defourny, 2010. Automated image-to-map discrepancy detection using iterative trimming, *Photogrammetric Engineering & Remote Sensing*, 76(2):173–181.

Radoux, J., C. Lamarche, E. Van Bogaert, S. Bontemps, C. Brockmann, and P. Defourny, 2014. Automated training sample extraction for global land cover mapping, *Remote Sensing*, 6(5):3965–3987.

Sarma, A., and D.D. Palmer, 2004. Context-based speech recognition error detection and correction, *Proceedings of HLT-NAACL 2004: Short Papers*, Association for Computational Linguistics, pp. 85–88.

Schistad Solberg, A. H.,T. Taxt, and A.K. Jain, 1996. A Markov random field model for classification of multisource satellite imagery, *IEEE Transactions on Geoscience and Remote Sensing* 34(1):100–113.

Subudhi, B.N., F. Bovolo, A. Ghosh, and L. Bruzzone, 2014. Spatio-contextual fuzzy clustering with Markov random field model for change detection in remotely sensed images, *Optics and Laser Technology*, 57(2014):284–292.

Wegner, J., F. Rottensteiner, M. Gerke, and G. Sohn, 2015. The ISPRS labelling challenge, URL: *http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html*, last date accessed: 15 March 2018.

SUSTAININGMEMBERS

Aerial Services, Inc.
Cedar Falls, Iowa
www.AerialServicesInc.com
Member Since: 5/2001

ASD Inc.,
a PANalytical Company
Boulder, Colorado
www.asdi.com
Member Since: 5/1995

Axis GeoSpatial, LLC
Easton, Maryland
www.axisgeospatial.com
Member Since: 1/2005

Ayres Associates, Inc.
Madison, Wisconsin
www.AyresAssociates.com
Member Since: 1/1953

Bohannan Huston, Inc.
Albuquerque, New Mexico
www.bhinc.com
Member Since: 11/1992

Cardinal Systems, LLC
Flagler Beach, Florida
www.cardinalsystems.net
Member Since: 1/2001

Certainty 3D LLC
Orlando, Florida
www.certainty3d.com
Member Since: 3/2011

DAT/EM Systems International
Anchorage, Alaska
www.datem.com
Member Since: 1/1974

Deimos Imaging
Boecillo - Valladolid, Spain
www.deimos-imaging.com
Member Since: 1/2014

Dewberry
Fairfax, Virginia
www.dewberry.com
Member Since: 1/1985

DigitalGlobe, Inc.
Longmont, Colorado
www.digitalglobe.com
Member Since: 6/1996

Environmental Research Incorporated
Linden, Virginia
www.eri.us.com
Member Since: 8/2008

Esri
Redlands, California
www.esri.com
Member Since: 1/1987

GeoBC
Victoria, Canada
www.geobc.gov.bc.ca
Member Since: 12/2008

GeoCue Group
Madison, Alabama
info@geocue.com
Member Since: 10/2003

Geomni, Inc.
Lehi, Utah
Geomni.net/psm
Member Since: 03/2018

GeoWing Mapping Inc.
Oakland, California
www.geowingmapping.com
Member Since: 1/2017

Global Science & Technology, Inc.
Greenbelt, Maryland
www.gst.com
Member Since: 10/2010

GPI Geospatial Inc.
formerly Aerial Cartographics of America, Inc. (ACA)
Orlando, Florida
www.aca-net.com
Member Since: 10/1994

GRW Aerial Surveys, Inc.
Lexington, Kentucky
www.grwinc.com
Member Since: 1/1985

Harris Corporation
Melbourne, Florida
www.harris.com
Member Since: 06/2008

Hexagon Geospatial
Norcross, Georgia
www.hexagongeospatial.com
Member Since: 4/2015

Keystone Aerial Surveys, Inc.
Philadelphia, Pennsylvania
www.keystoneaerialsurveys.com
Member Since: 1/1985

Kucera International
Willoughby, Ohio
www.kucerainternational.com
Member Since: 1/1992

Lead'Air, Inc.
Kissimmee, Florida
www.trackair.com
Member Since: 5/2001

LizardTech
Portland, Oregon
www.lizardtech.com
Member Since: 9/1997

Martinez Geospatial, Inc. (MTZ)
Eagan, Minnesota
www.mtzgeo.com
Member Since: 1/1979

MDA Information Systems LLC
Gaithersburg, Maryland
www.mdaus.com
Member Since: 1/1983

Merrick & Company
Greenwood Village, Colorado
www.merrick.com/gis
Member Since: 4/1995

Microsoft UltraCam Team (Vexcel Imaging, GmbH)
Graz, Austria
www.microsoft.com/ultracam
Member Since: 6/2001

Michael Baker Jr., Inc.
Beaver, Pennsylvania
www.mbakercorp.com
Member Since: 1/1950

Miller Creek Aerial Mapping, LLC
Seattle Washington
www.mcamaps.com
Member Since: 12/14

Office of Surface Mining
Denver, Colorado
www.tips.osmre.gov
Member Since: 8/2017

PCI Geomatics
Richmond Hill, Ontario, Canada
www.pcigeomatics.com
Member Since: 1/1989

PixElement
Columbus, Ohio
http://pixelement.com/
Member Since: 2/2017

Quantum Spatial, Inc.
Sheboygan Falls, Wisconsin
www.quantumspatial.com
Member Since 1/1974

Robinson Aerial Survey, Inc. (RAS)
Hackettstown, New Jersey
www.robinsonaerial.com
Member Since: 1/1954

Routescene, Inc.
Durango, Colorado
www.routescene.com/
Member Since: 12/2007

Sanborn Map Company
Colorado Springs, Colorado
www.sanborn.com
Member Since: 10/1984

Sidwell Company
St. Charles, Illinois
www.sidwellco.com
Member Since: 11/1992

SkyIMD - Imaging Mapping & Data
Richmond, California
www.skyimd.com
Member Since: 1/2017

Surveying And Mapping, LLC (SAM)
Austin, Texas
www.sam.biz
Member Since: 12/2005

Teledyne Optech
Toronto, Canada
www.teledyneoptech.com
Member Since: 1/1999

Terra Remote Sensing (USA) Inc.
Bellevue, Washington
www.terraremote.com
Member Since: 10/2016

The Airborne Sensing Corporation
Toronto, Canada
www.airsensing.com
Member Since: 7/2003

Towill, Inc.
San Francisco, California
www.towill.com
Member Since: 1/1952

University of Twente/Faculty ITC
Enschede, Netherlands
www.itc.nl
Member Since: 1/1992

U.S. Geological Survey
Reston, Virginia
www.usgs.gov
Member Since: 4/2002

Unmanned Experts, Inc
Denver, Colorado
www.unmannedexperts.com
Member Since: 4/2016

The Virginia Department of Transportation
Richmond, Virginia
Member Since: 2/2017

Woolpert LLP
Dayton, Ohio
www.woolpert.com
Member Since: 1/1985

Archetypal Analysis for Sparse Representation-Based Hyperspectral Sub-pixel Quantification

Lukas Drees, Ribana Roscher, and Susanne Wenzel

Abstract

The estimation of land cover fractions from remote sensing images is a frequently used indicator of the environmental quality. This paper focuses on the quantification of land cover fractions in an urban area of Berlin, Germany, using simulated hyperspectral EnMAP *data with a spatial resolution of 30 m × 30 m. We use constrained sparse representation, where each pixel with unknown surface characteristics is expressed by a weighted linear combination of elementary spectra with known land cover class. We automatically determine the elementary spectra from image reference data using archetypal analysis by simplex volume maximization, and combine it with reversible jump Markov chain Monte Carlo method. In our experiments, the estimation of the automatically derived elementary spectra is compared to the estimation obtained by a manually designed spectral library by means of reconstruction error, mean absolute error of the fraction estimates, sum of fractions, R^2, and the number of used elementary spectra. The experiments show that a collection of archetypes can be an adequate and efficient alternative to the manually designed spectral library with respect to the mentioned criteria.*

Introduction

The estimation of the degree of imperviousness as an indicator of environmental quality is subject of current research towards a time and cost efficient monitoring of urban areas [1]. Due to increasing land consumption in cities in the recent years, which has negative effects on the natural *water* cycle, the monitoring of land use in those areas is important [2].

Remote sensing data, such as imaging spectroscopy, builds a valuable basis to comprehensively map urban areas and quantify the imperviousness based on the spectral information (e.g., [3], [4]). Especially, hyperspectral imagery is a suitable source for mapping of such areas, because it offers a high spectral separability of different materials. However, generally, the temporal and spatial resolution is limited in comparison to sensors with lower spectral resolution. These limitations are partially overcome with the launch of missions such as Environmental Mapping and Analysis Program (EnMAP), which increases the availability of hyperspectral data and the temporal resolution [5]. Nevertheless, due to its spatial resolution, the provided data is mainly characterized by spectrally mixed pixels, which demands sophisticated sub-pixel quantification approaches in order to estimate the fraction of various land cover classes in each pixel.

In this context, several approaches have been developed comprising regression approaches [6], [7], probabilistic classification methods [8], [9], [10], and the usage of spectral libraries for spectral mixture analysis [11], [12]. An overview of a wide variety of unmixing approaches can, for example, be found in [13]. While the latter approach needs a spectral library containing elementary spectra of known materials, the first two approaches also require mixed spectra for learning an appropriate model. These mixed spectra can be derived from the image using information about known mixed pixels, or from synthetically mixed pixel, as it has been presented in [8] and [6].

When using spectral libraries, a critical step is the extraction of the elementary spectra. A manual extraction is time-consuming and requires human expert-knowledge and therefore, automatic extraction techniques have been an active field of research during the past decade (e.g., [13], [14]). Most of the algorithms rely on the assumption that the elementary spectra lie on a convex hull or a convex polytope enclosing the data distribution (e.g., [15], [16]). Based on this assumption, all data samples can be reconstructed by a non-negative linear combination of the elementary spectra. A promising approach from this group is the so-called archetypal analysis, which searches for extreme points (also known as archetypes) in the data distribution (e.g., [17], [18], [19], [20]). Archetypal analysis has already been successfully applied in the field of sport analytics [21], plant phenotyping [22], or text analysis [23]. A valuable extension to archetypal analysis is presented by [24], in which extreme points are extracted in the kernel space, enabling an efficient nonlinear unmixing. Besides the actual determination of elementary spectra, other challenges exist which need to be tackled: The number of elementary spectra is unknown beforehand and thus, a suitable amount of spectra needs to be extracted to make the set representative enough, but also small enough to keep the sub-pixel quantification robust and efficient. Moreover, many extraction techniques depend on the initialization and thus, a strategy needs to be defined to ensure a stable result (e.g., [25], [26]).

In this paper, we address the challenge of automatically finding a representative set of elementary spectra, including the automatic determination of the number of elements, for sub-pixel quantification. We perform the sub-pixel quantification using a freely available simulated EnMAP scene (Figure 1) of an urban area in Berlin, Germany [27][1], aiming at the estimation of a fraction map containing the classes *impervious surface, vegetation, soil, and water*. In order to determine the class fractions, we use sparse representation with non-negativity, L_0-sparsity and sum-to-one constraint. We exploit archetypal analysis to extract elementary spectra in a fully automatic and unsupervised way. Moreover, we perform archetypal analysis by simplex volume maximization (SiVM), which states an efficient selection method [28], [29].

The main contribution of this work is the combination of archetypal analysis with reversible jump Markov chain Monte Carlo method (rjMCMC, [30]) to obtain a spectral library of high representational power, yet a small number of elementary spectra. Using this approach, we are able to determine the

University of Bonn, Bonn, Germany
(ribana.roscher@uni-bonn.de).

1. http://pmd.gfz-potsdam.de/enmap/showshort.php?id=escidoc:1480925

Photogrammetric Engineering & Remote Sensing
Vol. 84, No. 5, May 2018, pp. 279–286.
0099-1112/18/279–286

doi: 10.14358/PERS.84.5.279

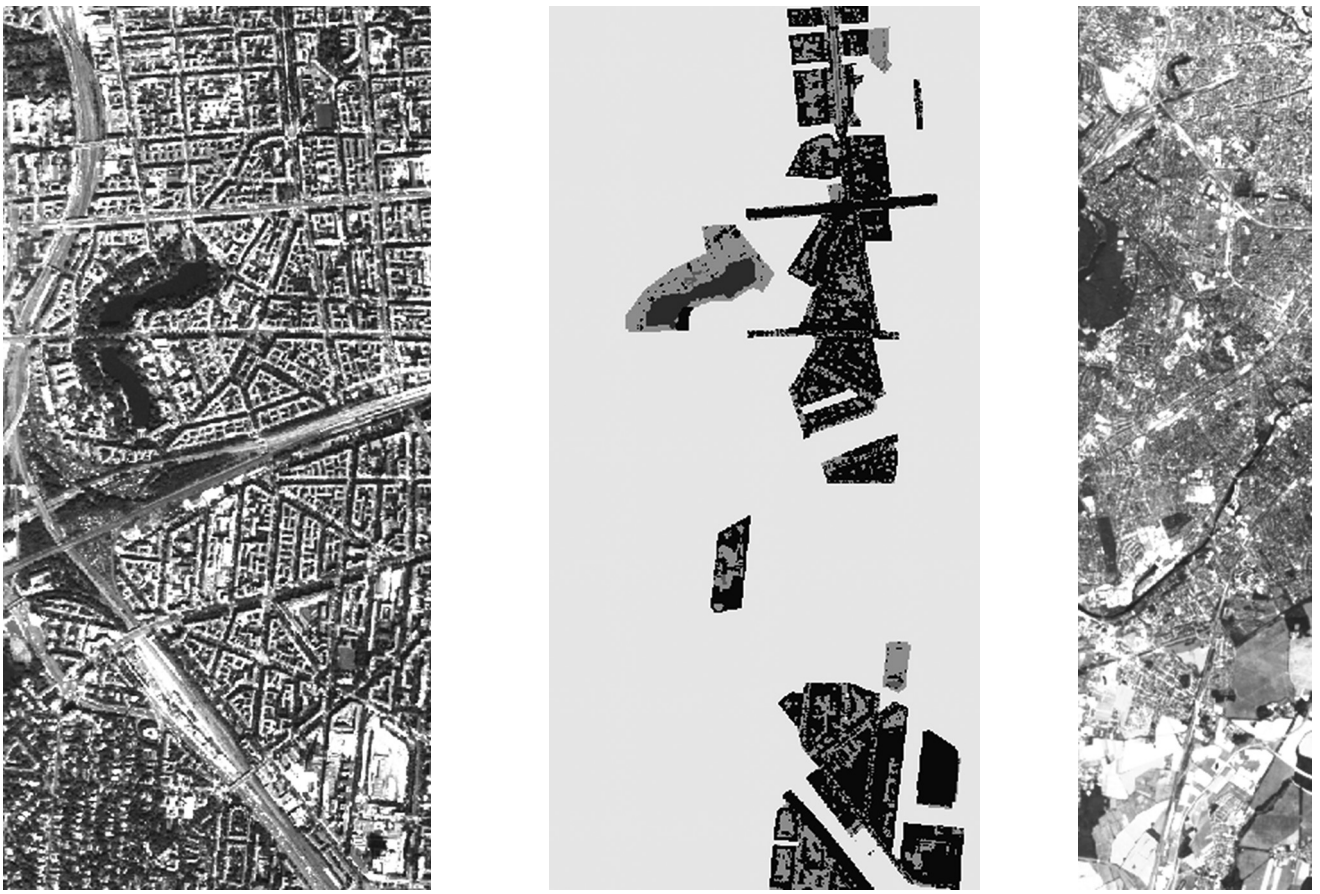


Figure 1. Berlin-Urban-Gradient dataset: (a) Part of HyMap image data visualized as RGB-image with the wavelengths R = 640 nm, G = 540 nm and B = 450 nm, (b) Reference image with four classes *impervious surface, vegetation, soil* and *water*, and (c) Simulated EnMAP data visualized with the wavelengths as specified above (bands: RGB = 11,5,1)

best set of spectral library elements regarding a pre-defined criteria, as well as the number of elements. Moreover, the approach of [24] is applied to select the archetypes in kernel space, resulting in archetypes lying on the concave hull of the data distribution in the original feature space. Our experiments confirm that these archetypes are more suitable for sub-pixel quantification than archetypes extracted from the convex hull exclusively. To illustrate the usefulness of our proposed approach, we analyze various kinds of automatically derived spectral libraries and compare them to manually designed spectral libraries. Our presented approach is flexible regarding the chosen estimation technique, such that the constrained sparse representation can be replaced by other approaches commonly used in the unmixing community [13], or other constraints such as sparsity induced by L_1-norm [31]. As such, archetypal analysis can be replaced by, e.g., endmember extraction methods presented in [13], and deliver the input for the rjMCMC method.

Data

Our studies are performed using the Berlin-Urban-Gradient dataset [27], illustrated in Figure 1. The dataset consists of two hyperspectral images of different spatial resolution, two simulated EnMAP scenes of different spectral resolution, a manually designed spectral library, reference land cover information, and reference fractions for evaluation, which are explained in more detail in the following paragraphs.

Table 1. Composition of 75 spectra in the manually designed spectral library (*LibCom*).

Imp. surface	*Vegetation*	*Soil*	*Water*
39	31	4	1

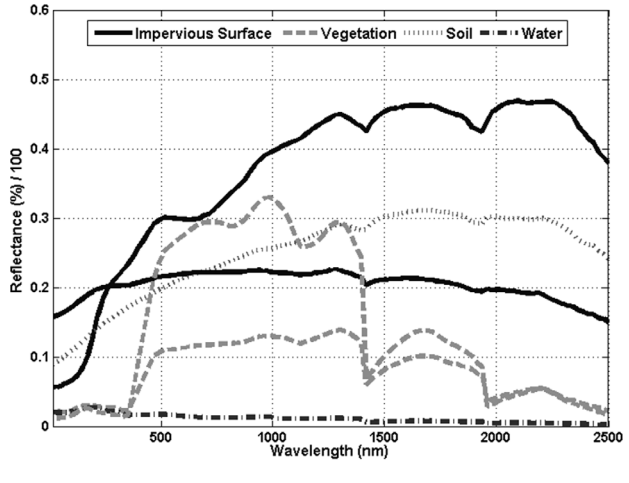


Figure 2. Significant spectra of the spectral library (*Lib*). Asphalt and red clay tile roof in the class *imp. surface, grass* and tree for *vegetation*, bare *soil,* and natural *water*.

HyMap Data

The dataset contains two hyperspectral images, one with a spatial resolution of 3.6 m and one with a spatial resolution of 9 m, whereby we use the higher resolution image in our work. Both images were acquired with the Hyperspectral Mapper (HyMap), and serve as basis for the extraction of the archetypes. It offers a high variety of urban land use and land cover patterns in the study site Southwest of Berlin, Germany. The external conditions were a cloudless sky at solar noon and a minimal possible altitude (maximal altitude according to the second lower resolution image). The HyMap image consists of 126 spectral bands; however, 15 noisy bands were removed resulting in 111 bands used for this work. Moreover, other preprocessing steps were performed, as encompassed system correction [32], atmospheric correction and parametric geocoding [33]. The observed wavelengths range from 0.45 µm to 2.5 µm, showing a high spectral information diversity, which enables a detailed analysis of the urban structure. In addition to the 3.6 m-resolution HyMap image, a manually-derived reference image containing 112.690 reference pixel is provided with valuable spectra, each labeled with one of the four land cover classes: *impervious surface, vegetation, soil,* and *water*. The reference information is manually obtained by using digital orthophotos and cadastral data. For our experiments, the set of elementary spectra is extracted from this data to obtain the class membership of the elementary spectra, however, the extraction can be performed on unlabeled data with limited human user interaction.

Simulated EnMAP Data

EnMAP is German hyperspectral satellite mission, which will start not earlier than 2018, with a focus on Earth environmental observations in a global scale. Based on the HyMap data, an EnMAP scene was simulated using EnMAP end-to-end simulation tool (EeteS, [34]) with two different spectral resolutions (111 and 244 bands). Just as the HyMap data, both EnMAP images have a spectral resolution ranging from 0.45 µm to 2.5 µm, where we use 111 bands for a better comparability. The spatial resolution of 30 m is lower than the resolution of the HyMap scene, and thus the mixing of land cover classes are more apparent. For evaluation purposes, 1495 EnMAP pixels were obtained from the simulation tool, containing the fractions of the land cover classes ranging from 0 to 100 %.

Manually Designed Spectral Library

The spectral library is a manually designed collection of 75 pure spectra obtained from the HyMap image. The spectra contain different land cover classes with 39 *impervious surface* spectra (different types of roof, pavement, tartan, pool *water*), 31 *vegetation* spectra (grass, tree), 4 *soil* spectra (uncovered ground, sand) and one natural *water* spectrum. All together, it is a balanced library for urban structures with 39 impervious and 36 pervious spectra. All spectra in the library can be assigned to a hierarchical urban classification scheme, which was developed by [35]. First, an initial collection of 300 spectra was selected by expert knowledge. Afterwards a two-step filtering was performed, in which the variability between the spectra is maximized in consideration of spectral variability of materials, brightness and shading effects. Moreover, in an iterative process those spectra for the final subset were selected that best describes the spectral diversity in a specific neighborhood. More details can be found in [27].

Methods

The following section describes the sub-pixel quantification using sparse representation and archetypal analysis. Archetypal analysis determines the extreme points of the data distribution, which are used within the sparse representation approach to estimate the fractions of land cover classes in each pixel. We have given a ($M\times N$)-dimensional data matrix $\boldsymbol{X}$, in which N is the number of M-dimensional reference pixels $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]$ with given land cover class c_n. Moreover, we have given a test set of 1,495 data samples ${}^T\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T]$ with known land cover fractions ${}^T\boldsymbol{f}_t$ for evaluation purposes, as presented in the Data Section.

Sparse Representation

In order to determine the sub-pixel fractions, we use sparse representation with non-negative least squares optimization. In terms of sparse coding, a sample x is represented by a weighted linear combination of a few elements taken from a ($M\times D$)-dimensional dictionary $\boldsymbol{D}$, such that $\boldsymbol{x}=\boldsymbol{D}\boldsymbol{\alpha}+\boldsymbol{\gamma}$ with $\|\gamma_2\|$ being the reconstruction error. The dictionary $\boldsymbol{D}=[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d, \ldots \boldsymbol{x}_D]$ contains elementary spectra, such that this approach is identical to the linear mixing model. The coefficient vector, comprising the activations, is given by $\boldsymbol{\alpha}$. The activations are interpreted as class fractions for sub-pixel quantification. The optimization problem for the determination of optimal $\hat{\alpha}$ is given by

$$\hat{\boldsymbol{\alpha}} = \operatorname{argmin} \|\boldsymbol{D}\boldsymbol{\alpha} - \boldsymbol{x}\|_2, \tag{1}$$

$$\textit{subject to } \boldsymbol{\alpha} \geq 0,\ \textstyle\sum \boldsymbol{\alpha}_d = 1,\ \|\boldsymbol{\alpha}\|_0 \leq W \tag{2}$$

where the terms in Equation 2 are the non-negativity constraint, the sum-to-one constraint, and the sparsity constraint. Generally, non-negativity alone leads to a sparse solution, however, in order to ensure a strict fulfillment of the sparsity constraint, we use a backward selection procedure in which the activations with the smallest values are set to zero in a greedy manner.

Archetypal Analysis

Archetypal analysis is a suitable method to determine the elements of $\boldsymbol{D}$, where each archetype serves as one dictionary element. The extraction of the archetypes, collected in a ($M\times K$) -dimensional matrix of $\boldsymbol{A} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_k, \ldots, \boldsymbol{\alpha}_K]$, $k = 1,\ldots,K$, is carried out by SiVM, which is an efficient method to determine the archetypes of the data distribution. This approach is aiming on an approximation of the convex hull, where all archetypes are located on it.

In order to find the first archetype, the approach is initialized with a random or pre-defined vector $\boldsymbol{\alpha}_0$. The pixel with maximal distance to $\boldsymbol{\alpha}_0$, is defined as first archetype $\boldsymbol{\alpha}_1$, defined by

$$\boldsymbol{\alpha}_1 = \arg\max_n \boldsymbol{d}(\boldsymbol{\alpha}_0, \boldsymbol{x}_n), \tag{3}$$

with $d(.\ ,.)$ being the Euclidean distance function between the spectral features of the archetype a_0 and the pixel x_n. Further archetypes are specified sequentially, such that the volume of the simplex becomes maximized with each additional archetype. Since the volume operation is too computational intense, instead the archetypes are selected to have maximum distance to all previously detected ones, using:

$$\boldsymbol{\alpha}_M = \arg\max_n \sum_k d(\boldsymbol{\alpha}_k, \boldsymbol{x}_n). \tag{4}$$

The stopping criterion is generally chosen to be the number of archetypes.

We further use the approach of [24] and transform $\boldsymbol{X}$ into kernel space using a Gaussian radial basis function kernel with a hyperparameter σ describing the width of the Gaussian kernel. In this way, archetypes are selected which lie on the concave hull rather than the convex hull.

The disadvantage of archetypal analysis is that the final set depends on the initialization point, and as a result, there is no unique solution to the final set. Especially, if the number of archetypes in the dictionary is low, various solutions lead to significantly different accuracies. Moreover, the number of archetypes

is generally not know beforehand, and depends on the number of informative dimensions and the variability of the data.

Reversible Jump Markov Chain Monte Carlo Method

To overcome this problem, we propose to use an optimization procedure to find the best set of archetypes from a large set of pre-selected ones, called the initial set. Under the assumption that a suitable archetypal set is able to represent the test set ${}^{\mathrm{T}}X$ with a low reconstruction error, our task is to find the set of archetypes $D = A$ which minimizes the energy

$$\mathfrak{V}(D) = \|\Upsilon\|_2 \tag{5}$$

where Υ is obtained by using the current set of archetypes as dictionary D. The energy $\mathfrak{V}$ is a complex function with rough landscape and unknown dimensionality due to the unknown number of archetypes. Therefore, we optimize with rjMCMC coupled with simulated annealing to find the global optimum. Introducing the temperature parameter R, the optimizer is given by:

$$\hat{D} = \arg\min_{D} \frac{\mathfrak{V}(D)}{R_k}, \quad \lim_{k\to\infty} R_k = 0 \tag{6}$$

While MCMC is dedicated to sample from complex functions, simulated annealing allows to make a point estimate of its global optimum. Using simulated annealing we create a Markov chain, such that the samples explore the whole state space in the beginning and gradually concentrate around the global optimum of the energy function $\mathfrak{V}$. In this way we avoid trapping into local optima, as it is usually the case for greedy algorithms. We use the so-called birth and death algorithm [36] to sample from a restricted sample space of possible sets of archetypes, which turns out to be a special type of Green's rjMCMC sampler [30]. The sample space is restricted by a Poisson prior on the expected number of archetypes, as presented in [37]. We further introduce an upper bound on the the number of selected archetypes.

Experiments

Experimental Setup

In our experiments, the simulated EnMAP data is reconstructed by sparse representation with the before mentioned constraints using the following libraries:

1. the manually designed spectral library (*LibCom*),
2. an optimized set of the manually designed spectral library *LibRed* (using rjMCMC),
3. a library containing 40 extracted archetypes in the original feature space initialized by the mean vector (*AA-M-Lin*),
4. a library containing 75 extracted archetypes in the kernel space initialized by the mean vector (AA-M),
5. a library containing 75 extracted archetypes in the kernel space initialized by a random vector (*AA-Rand*),
6. a library containing an accumulated set of multiple single sets of archetypes, obtained in the kernel space with random initializations (AA-Full), and
7. an optimized set of AA-Full denoted by AA-Opt (using rjMCMC).

We compare our results to the regression and classification methods presented in [8], namely support vector machine (SVM), import vector machine (IVM), support vector regression (SVR), and multi-output support vector regression (MSVR). The aim of the experiments is to show the suitability of the spectral libraries for sub-pixel quantification. Moreover, the goal is to show that the set of automatically derived spectral libraries using archetypal analysis and rjMCMC achieve similar results than the manually designed spectral library. The sub-pixel quantification is evaluated with the given reference mixing fractions. Additionally, the number of used archetypes is evaluated and discussed. We run all experiments including a random component 10 times and report the average result and standard deviation.

As preprocessing step, the dataset is outlier-cleaned using the local outlier factor approach [38], using 10 neighbors and a quantile of 0.95 as threshold on the pairwise Euclidean distances. The width of the Gaussian kernel σ is chosen to be 0.5 of the average standard deviation over all bands. In order to choose a suitable sparsity value W, we use Elbow method to analyze the reconstruction error of the test set obtained by the manually designed library as well as the automatically derived library containing 75 archetypes. Figure 3 shows that the reconstruction error will not significantly decrease for $W > 7$, and thus we choose this value as upper bound for the sparsity constraint. For rjMCMC we choose the prior on the expected number of archetypes to be 75 when using all labeled data, and 40 when using the manually designed spectral library. The upper limit of archetypes is chosen to be 150.

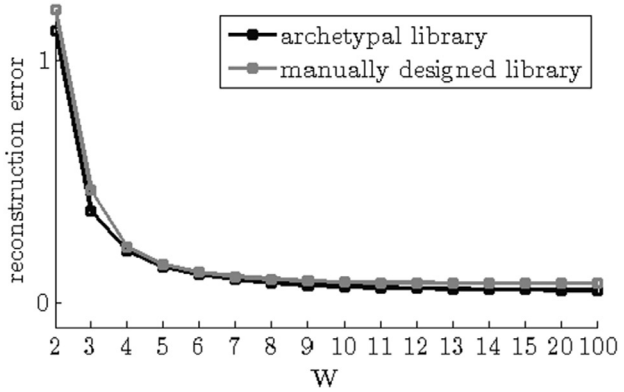


Figure 3. Reconstruction error evaluated on the test set for different sparsity values W.

Evaluation

We use several statistics for evaluation. First, the reconstruction error of the sparse representation is provided, indicating the ability of the elementary spectra to represent the EnMAP pixels. Moreover, the result of the sparse representation is evaluated by a comparison between reference and actual estimated fraction coefficients. This is done by means of the overall and class-wise mean absolute error (MAE) between reference and actual coefficients.

$$MAE = \frac{1}{T}\sum_{t}^{T}\left|\left({}^{\mathrm{T}}f_{\mathrm{t}} - {}^{\mathrm{I}}f_{\mathrm{t}}\right)\right| \tag{7}$$

with ${}^{T}f_{\mathrm{t}}$ being the reference fractions, ${}^{I}f_{\mathrm{t}}$ being the estimated fractions, and T being the number of evaluated pixels. We obtain ${}^{I}f_{\mathrm{t}}$ by summing up all coefficients belonging to elementary spectra of the same class. The smaller the MAE, the better the reconstruction. Moreover, the root mean square error (*RMSE*) is provided, which is defined by:

$$RMSE = \frac{1}{T}\sum_{t}^{T}\sqrt{\left({}^{\mathrm{T}}f_{\mathrm{t}} - {}^{\mathrm{I}}f_{\mathrm{t}}\right)^2} \tag{8}$$

Besides this, we provide the coefficient of determination (R^2) for each class. It is calculated as the squared correlation coefficient between ${}^{T}f_{\mathrm{t}}$ and ${}^{I}f_{\mathrm{t}}$.

Results and Discussion

Table 2 presents the results of the EnMAP sub-pixel quantification. In the first block, the results of [8] are shown for

comparability with regression and classification results. The middle block shows the results obtained by the full and reduced manually designed spectral library. The bottom part of the table presents the results obtained by archetypal analysis.

Overall, the MAE values show that all spectral libraries which are obtained from the full manually designed library X_{lib} achieve similar and satisfactory results. It can be seen that in comparison to the results of the regression and classification approaches presented in [8], sparse representation with the mentioned constraints show equivalent results. Both spectral libraries have a small average MAE <9 percent, which is also obtained by IVM and *MSVR*. Remarkably, *LibCom* and *LibRed* show similar results, however, with *LibRed* having much fewer elements. *LibCom* has slightly lower class-wise MAE for the classes *impervious surface*, *vegetation*, and *soil*, where *LibRed* obtains better results for *water*. This indicates that the presented rjMCMC method is able to find a suitable subset of elementary spectra having nearly the same representational power than the manually designed library with all elements. Moreover, it can be stated from this result that the full spectral library contains redundant elementary spectra which can be discarded for sub-pixel quantification.

These findings are underlined by Figure 4, which shows the scatter plots representing the 1,495 class-wise fraction estimates opposed to the reference fractions. It can be observed that *soil* and *water* have a high proportion of 0 percent reference fractions and lesser >0 percent fractions than *impervious surface* and *vegetation*. The *impervious surface* scatter plot show that for *LibCom* the two lines intersect nearly at the 15 percent point on the x-axis. This means that the estimated fractions are usually too small for all pixels which have a degree of impervious surfaces over 15 percent, and otherwise too high for smaller values than the intersection point. In comparison to *LibRed*, the reference values are also mostly underestimated, but slightly better than *LibCom*. The *soil* scatter plots have large discrepancies between the true 1:1 line (dashed line) and the estimated least-squares regression line (solid line). Besides this, there is a high amount of 0% fractions, which are estimated with up to 40% when using the *LibCom* spectra and 30 percent fractions when using *LibRed* spectra. Furthermore, all pixels with *soil*-fractions are clearly underestimated equally. This is also the case for the regression and classification approaches presented in [8]. We assume that the small number of elementary spectra is insufficient for the spectral diversity of *soil* pixels. Finally, the described observation with high estimated 0 percent fractions also occurs in the Water class with values over 50 percent.

Figure 5 illustrates the frequency of the usage of each spectrum for reconstruction in order to determine the fractions of 1,493 EnMAP pixels. The results obtained by *LibCom* are shown on the left and the results for *LibRed* on the right side. Moreover, the lower bright part of each bar indicates the sum over all fractions from which the total proportion of each land cover class in the study area can be derived. *LibCom* shows that there are significant elementary spectra, e.g., 18, 19 (*impervious surface*), 45, 48, 52, 54 (*vegetation*), and 75 (*water*), which show a high fraction in the reconstruction. In the other side, some elementary spectra in this plot are infrequently used. It can be observed that there is generally no dependence between the height of a bar, i.e., the number of non-zero estimated fractions for the spectra, and its overall fraction's sum for all pixels. A striking example is spectrum 19. Spectrum 19 in the *LibCom* library is used for more than 500 pixel reconstructions, but its fraction's sum is lower in comparison to, e.g., spectrum 18, which is used more infrequently. However, *LibRed* has a more balanced composition of elementary spectra with fewer bar heights, which are close to zero. Especially noticeable are the differences in the bar for the class *water* between *LibCom* and *LibRed*. A *water* spectrum has the special characteristic that there are only small reflectance values over all bands, so that it acts almost as a linear factor. Because of this, it is well suited to support all reconstructions, resulting in a high overall fractions' sum. Its value in the *LibCom* is high with a share in over 800 reconstructions, in comparison to *LibRed* which shows a bar height of just over 600 and a lower fractions' sum. We assume the reason for this is the smaller number of elementary spectra, where oftentimes a reconstruction with *water* spectra is helpful regardless of the presence of *water* in the pixel.

The results in Table 2 obtained from various spectral libraries based on the labeled data set X are more variable, indicating their dependency of a proper choice of hyperparameters such as the number of archetypes and the selection approach. For example, the number of selected archetypes in set *AA–M–Lin* is 40, which is smaller in comparison to *AA–M*. Both sets are initialized with the mean vector of all samples in X, where the first set is extracted in the original feature space and the latter one in the kernel space. We observed that a selection

Table 2. Evaluation results of EnMAP sub-pixel quantification. Overall (Ø) and class-wise *MAE* (in brackets computed without zero-reference fractions) and overall *RMSE*. Sparse representation-based estimations with various obtained spectral libraries (for explanations, see the Experimental Setup Section) are compared to regression (SVR, MSVR) and classification methods (IVM, SVM). The standard deviation values are indicated by ±.

Data	Approach	Amount of elements	Class-wise MAE [%]				Ø MAE [%]	Ø RMSE [%]
			Imp. surface	*Vegetation*	*Soil*	*Water*		
X_{lib}	SVM	75	14.81 (15.22)	16.97 (15.97)	04.84 (19.01)	04.25 (66.33)	10.22	17.36
X_{lib}	IVM	75	15.21 (15.91)	15.31 (15.59)	02.09 (20.98)	01.98 (35.51)	08.65	15.63
X_{lib}	SVR	75	11.73 (11.51)	12.20 (11.92)	08.36 (12.78)	07.95 (07.68)	10.06	13.74
X_{lib}	MSVR	75	11.33 (11.45)	10.99 (11.49)	02.17 (13.45)	03.19 (09.70)	06.92	11.32
X_{lib}	LibCom	75	13.00 (13.79)	09.61 (09.87)	02.01 (14.37)	07.96(07.88)	08.15	12.66
X_{lib}	LibRed	44.4	14.00 (14.86)	10.58 (10.85)	02.23 (15.07)	07.46 (09.07)	08.57	13.68
		± 9.5	± 1.45 (1.57)	± 0.35 (0.39)	± 0.64 (0.61)	± 2.00 (1.71)	± 0.96	± 1.22
X	AA-M-Lin	40	25.09 (26.23)	15.05 (15.70)	01.62 (20.38)	33.00 (09.65)	18.69	23.65
X	AA-M	75	25.50 (27.00)	22.44 (22.93)	01.80 (02.04)	08.87 (20.39)	14.65	20.01
X	AA-Rand	75	21.05(21.58)	17.86 (18.27)	01.77 (15.69)	08.52 (26.29)	12.30	18.17
		± 0	±3.11(4.13)	±3.05(3.10)	±0.06(3.11)	±4.07(27.91)	±2.00	±1.22
X	AA-Full	142	49.06 (52.71)	32.50 (33.20)	02.38 (26.16)	2.84 (68.13)	21.70	31.45
X	AA-Opt	81.7	16.29 (16.21)	14.21 (14.70)	03.14 (24.09)	07.00 (09.73)	10.16	16.08
		± 13.6	± 0.69 (0.76)	± 0.65 (0.60)	± 0.60 (2.91)	± 1.50 (1.38)	± 0.51	± 0.47

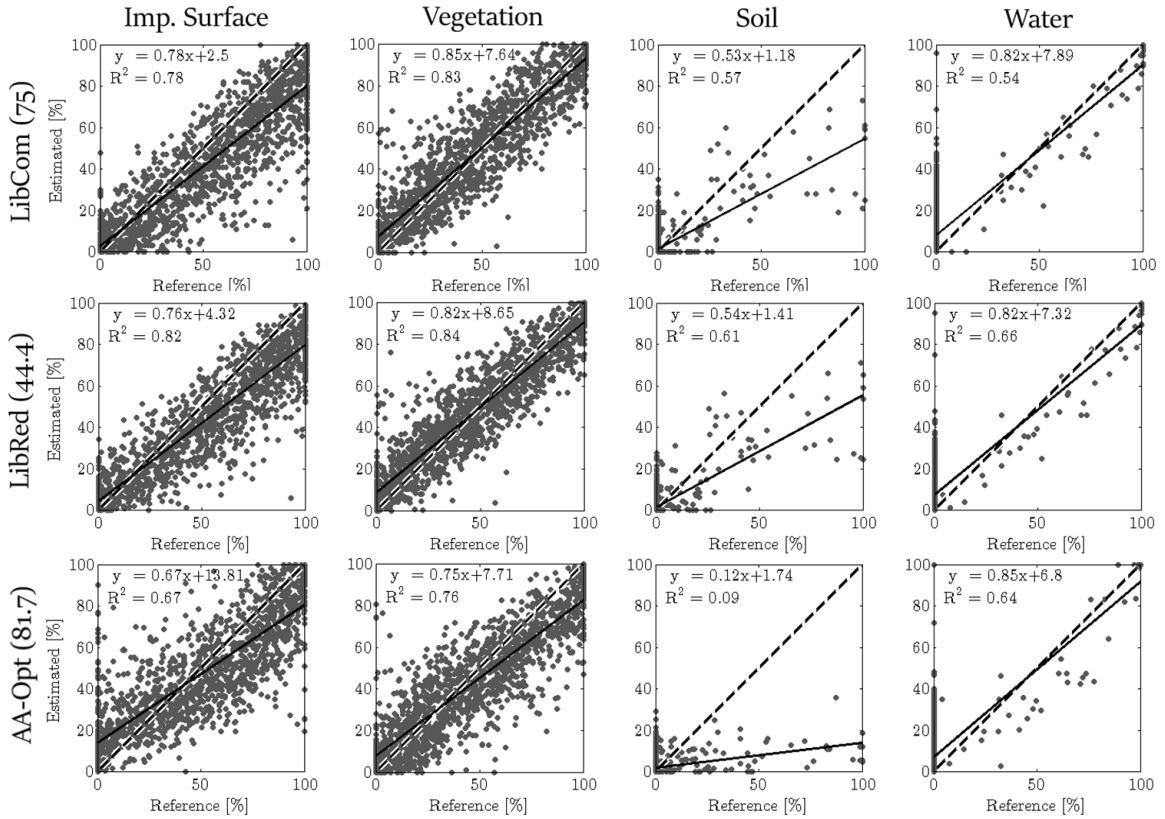
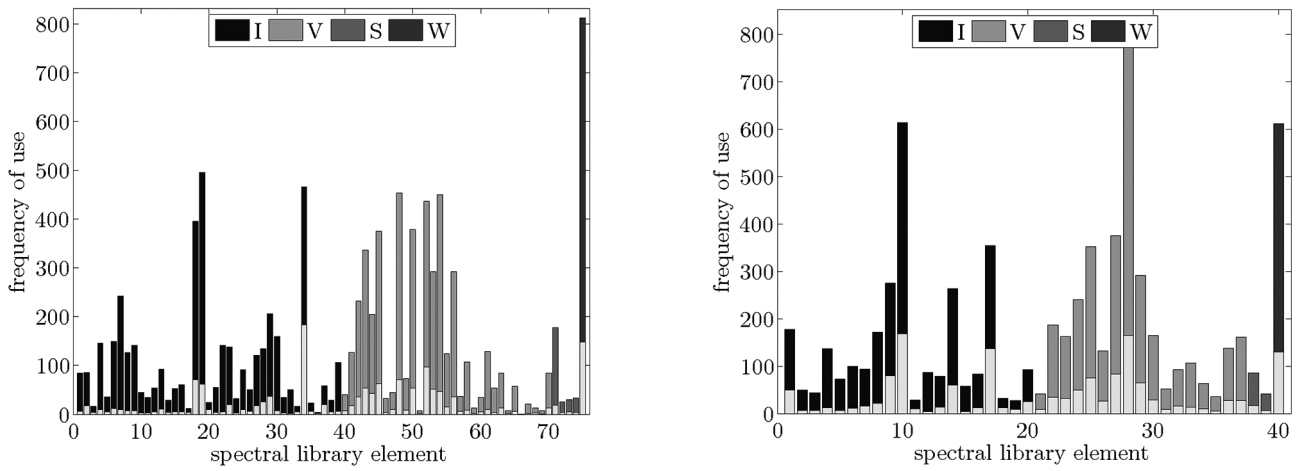


Figure 4. Scatter plots obtained from *LibCom* (top row), *LibRed* (middle row) and *AA-Opt* (bottom row) representing the class-wise fraction estimates of impervious surface, vegetation, soil and water, opposed to the reference fractions. A correct estimated pixel lies on the dashed line. The solid line represents the respective least-square regression line for the scatter points. The formula for the regression line and coefficient of determination (R^2) is also given in each plot.



(a) Full manually designed spectral library *LibCom* **(b) Reduced manually designed spectral library *LibRed***

Figure 5. The frequency of the usage of elementary spectra for the reconstruction of the reference EnMAP pixels. Different colors indicate the class membership of the reference pixels (I: *Imp. surface*, V: *Vegetation*, S: *Soil*, W: *Water*). The maximal possible height of a bar is 1,493, if the elementary spectrum is involved in the reconstruction of all reference EnMAP pixels. The lower bright part of each bar specify the sum of the fractions (i.e., the estimated coefficients) over all pixel.

of more than 40 archetypes in the set *AA–M–Lin* means that some archetypes lie in the center of the dataset X, since they have the maximal distance to the previously selected archetypes. The set *AA–M–Lin* consists of 25 *impervious surface* spectra, 12 *vegetation* spectra, 2 *soil* spectra and one *water* spectrum. The total number of selected archetypes in *AA-M* is chosen to be 75, as for the manually designed library *LibCom*. For all classes except the *water* class, the results obtained by *AA-M* are worse than these ones obtained by *AA-M-Lin*. Also the results for the randomly initialized archetypal sets *AA-Rand* show only slightly better results. Moreover, we observed that oftentimes the selected archetypal set with random initialization does not contain all land cover classes. Thus, as indicated by the results, a single set on selected archetypes is not suitable enough for an accurate sub-pixel quantification. Therefore, in order to create a higher diversity of archetypes,

various initializations are used for SiVM and accumulated into a larger set. The set AA-Full contains 142 different archetypes, however, the high amount of elementary spectra results in high MAE values. The best result based on $\boldsymbol{X}$ is obtained by *AA-Opt*, which is the reduced AA-Full set. This set has the best overall MAE and the class-wise MAE are in most cases comparable to the approaches based on $\boldsymbol{X}_{lib}$. The final number of elements in set *AA-Opt* is 81.7 on average, which is similar to the number of elements in the manually designed spectral library.

Figure 4 underlines these findings. For *impervious surface*, *vegetation* and *water* the results are similar for all three libraries. However, also for *soil* the set *AA-Opt* has large discrepancy between the true 1:1 line (dashed line) and the estimated least-squares regression line (solid line), which is more distinct than for *LibCom* and *LibRed*.

Conclusions

The quantification of sub-pixel fractions in remote sensing images is a relevant task to assess the environmental quality, for example, in urban areas. This paper presents a sub-pixel quantification of the urban area of Berlin, Germany, into four land cover classes *impervious surface*, *vegetation*, *soil*, and *water*, using a manually designed spectral library and various kinds of automatically derived spectral libraries. We perform the estimation of the fractions by using sparse representation with non-negativity, sparsity, and sum-to-one constraints. We use archetypal analysis by simplex volume maximization to automatically derive the elements for a library, and apply reversible jump Markov chain Monte Carlo method to find a small, yet representative set of suitable elementary spectra. The archetypes are extracted from HyMap data with given reference information for all land cover classes. As our experiments suggest, the extracted archetypes are suitable to serve as spectral library for sub-pixel quantification. Moreover, in contrast to a manually designed library, the automatically derived spectral library can be easily extracted with no or limited human user interaction, and the library is specifically adapted to the current image characteristics. In case no reference data is given, the interpretation can be done in a fast way by assigning land cover classes to the extracted archetypes using expert knowledge. The presented approach is flexible regarding the chosen estimation technique for the sub-pixel fractions, and can also be combined with spatial information, which may further increase the approximation ability and accuracy of the fraction estimates.

Acknowledgments

The authors would like to thank the reviewers for their valuable comments, and Akpona Okujeni, Sebastian van der Linden, and Patrick Hostert for providing the dataset. Moreover, the authors would like to thank Johannes Rosentreter and Andres Milioto for valuable discussions.

References

[1] Q. Weng, Remote sensing of impervious surfaces in the urban areas: Requirements, methods, and trends, *Remote Sensing of Environment*, vol. 117, pp. 34–49, 2012.

[2] G. Wessolek, Bodenüberformung und -versiegelung, *Handbuch der Bodenkunde*, 2001.

[3] X. Huang, Q. Lu, and L. Zhang, A multi-index learning approach for classification of high-resolution remotely sensed images over urban areas, *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 90, pp. 36–48, 2014.

[4] S. Roessner, K. Segl, M. Bochow, U. Heiden, W. Heldens, and H. Kaufmann, 2011. Potential of hyperspectral remote sensing for analyzing the urban environment, *Urban Remote Sensing: Monitoring, Synthesis and Modeling in the Urban Environment*, pp. 49–61.

[5] L. Guanter, H. Kaufmann, K. Segl, S. Foerster, C. Rogass, S. Chabrillat, T. Kuester, A. Hollstein, G. Rossner, C. Chlebek, 2015. The enmap spaceborne imaging spectroscopy mission for earth observation," *Remote Sensing*, vol. 7, no. 7, pp. 8830–8857.

[6] A. Okujeni, S. van der Linden, S. Suess, and P. Hostert, 2016. Ensemble learning from synthetically mixed training data for quantifying urban land cover with support vector regression, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2016.

[7] F. Priem, A. Okujeni, S. van der Linden, and F. Canters, 2016. Use of multispectral satellite imagery and hyperspectral endmember libraries for urban land cover mapping at the metropolitan scale," in *SPIE Remote Sensing*, International Society for Optics and Photonics.

[8] J. Rosentreter, R. Hagensieker, A. Okujeni, R. Roscher, and B. Waske, 2017. Sub-pixel mapping of urban areas using enmap data and multioutput support vector regression, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, in press.

[9] S. Suess, S. van der Linden, P. J. Leitao, A. Okujeni, B. Waske, and P. Hostert, 2014. Import vector machines for quantitative analysis of hyperspectral data, *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 2, pp. 449–453.

[10] J. Zhu and T. Hastie, Kernel logistic regression and the import vector machine, *Journal of Computational and Graphical Statistics*, 2012.

[11] B. Somers, G. P. Asner, L. Tits, and P. Coppin, 2011. Endmember variability in spectral mixture analysis: A review, *Remote Sensing of Environment*, vol. 115, no. 7, pp. 1603–1616.

[12] R. L. Powell, D. A. Roberts, P. E. Dennison, and L. L. Hess, 2007. Sub-pixel mapping of urban land cover using multiple endmember spectral mixture analysis: Manaus, Brazil, *Remote Sensing of Environment*, vol. 106, no. 2, pp. 253–267.

[13] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot, 2012. Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 2, pp. 354–379, 2012.

[14] M. A. Veganzones and M. Grana, 2008. Endmember extraction methods: A short review," *Proceedings of the International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, Springer, pp. 400–407.

[15] T.-H. Chan, W.-K. Ma, A. Ambikapathi, and C.-Y. Chi, 2011. A simplex volume maximization framework for hyperspectral endmember extraction, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 11, pp. 4177–4193.

[16] M. D. Craig, Minimum-volume transforms for remotely sensed data, 1994. *IEEE Transactions on Geoscience and Remote Sensing*, vol. 32, no. 3, pp. 542–552.

[17] C. Zhao, G. Zhao, and X. Jia, 2017. Hyperspectral image unmixing based on fast kernel archetypal analysis, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 1, pp. 331–346.

[18] G. Zhao, C. Zhao, and X. Jia, 2016. Multilayer unmixing for hyperspectral imagery with fast kernel archetypal analysis, *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 10, pp. 1532–1536.

[19] M. Mørup and L. K. Hansen, 2012. Archetypal analysis for machine learning and data mining, *Neurocomputing*, vol. 80, pp. 54–63.

[20] A. Cutler and L. Breiman, 1994. Archetypal analysis, *Technometrics*, vol. 36, pp. 338–347.

[21] S. Seth and M. J. Eugster, 2016. Probabilistic archetypal analysis, *Machine Learning*, vol. 102, no. 1, pp. 85–113.

[22] C. Römer, M. Wahabzada, A. Ballvora, F. Pinto, M. Rossini, C. Panigada, J. Behmann, J. Léon, C. Thurau, C. Bauckhage, 2012. Early drought stress detection in cereals: Simplex volume maximization for hyperspectral image analysis, *Functional Plant Biology*, vol. 39, no. 11, pp. 878–890.

[23] C. Seiler and K. Wohlrabe, 2013. Archetypal scientists, *Journal of Informetrics*, vol. 7, no. 2, pp. 345–356.

[24] R. Heylen, P. Scheunders, A. Rangarajan, and P. Gader, 2015. Nonlinear unmixing by using different metrics in a linear unmixing chain, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 6, pp. 2655–2664.

[25] Wang, H. Li, W. Liao, and X. Huang, 2016. Endmember initialization method for hyperspectral data unmixing, *Journal of Applied Remote Sensing*, 2009. vol. 10, no. 4, p. 042009.

[26] M. Zortea and A. Plaza, 2009. A quantitative and comparative analysis of different implementations of n-findr: A fast endmember extraction algorithm, *IEEE Geoscience and Remote Sensing Letters*, vol. 6, no. 4, pp. 787–791.

[27] A. Okujeni, S. van der Linden, and P. Hostert, 2016. Berlin-urban-gradient dataset 2009 - An enmap preparatory flight campaign (datasets), *GFZ Data Services*, 2016.

[28] C. Thurau, K. Kersting, and C. Bauckhage, 2010. Yes we can: Simplex volume maximization for descriptive web-scale matrix factorization, *Proceedings of the International Conference on Information and Knowledge Management*, ACM, pp. 1785–1788.

[29] R. Heylen, D. Burazerovi , and P. Scheunders, 2011. Non-linear spectral unmixing by geodesic simplex volume maximization, *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 3, pp. 534–542.

[30] P. J. Green, 1995. Reversible jump markov chain monte carlo computation and bayesian model determination, *Biometrika*, vol. 82, no. 4, pp. 711–732.

[31] Z. Zhang, Y. Xu, J. Yang, X. Li, and D. Zhang, 2015. A survey of sparse representation: algorithms and applications," *IEEE Access*, vol. 3, pp. 490–530.

[32] T. Cocks, R. Jenssen, A. Stewart, I. Wilson, and T. Shields, 1998. The hymap airborne hyperspectral sensor: The system, calibration and performance, in *Proceedings of the 1st EARSeL Workshop on Imaging Spectroscopy*, EARSeL, pp. 37–42.

[33] R. Richter and D. Schläpfer, 2002. Geo-atmospheric processing of airborne imaging spectrometry data. Part 2: atmospheric/topographic correction, *International Journal of Remote Sensing*, vol. 23, no. 13, pp. 2631–2649.

[34] K. Segl, L. Guanter, C. Rogass, T. Kuester, S. Roessner, H. Kaufmann, B. Sang, V. Mogulsky, and S. Hofer, 2012. Eetesthe enmap end-to-end simulation tool, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 2, pp. 522–530.

[35] U. Heiden, K. Segl, S. Roessner, and H. Kaufmann, 2007. Determination of robust spectral features for identification of urban surface materials in hyperspectral remote sensing data," *Remote Sensing of Environment*, vol. 111, no. 4, pp. 537–552.

[36] C. J. Geyer and J. Møller, 1994. Simulation Procedures and Likelihood Inference for Spatial Point Processes, *Scandinavian Journal of Statistics*, vol. 21, no. 4, pp. pp. 359–373.

[37] D. Bulatov, S. Wenzel, G. Häufel, and J. Meidow, 2017. Chain-wise generalization of road networks using model selection, *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. IV-1/W1, pp. 59–66.

[38] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, 2000. Lof: Identifying density-based local outliers, *ACM Sigmod Record*, vol. 29, pp. 93–104.

Classification of Aerial Photogrammetric 3D Point Clouds

C. Becker, E. Rosinskaya, N. Häni, E. d'Angelo, and C. Strecha

Abstract

We present a powerful method to extract per-point semantic class labels from aerial photogrammetry data. Labeling this kind of data is important for tasks such as environmental modeling, object classification, and scene understanding. Unlike previous point cloud classification methods that rely exclusively on geometric features, we show that incorporating color information yields a significant increase in accuracy in detecting semantic classes. We test our classification method on four real-world photogrammetry datasets that were generated with Pix4Dmapper, and with varying point densities. We show that off-the-shelf machine learning techniques coupled with our new features allow us to train highly accurate classifiers that generalize well to unseen data, processing point clouds containing 10 million points in less than three minutes on a desktop computer. We also demonstrate that our approach can be used to generate accurate Digital Terrain Models, outperforming approaches based on more simple heuristics such as Maximally Stable Extremal Regions.

Introduction

Extraction of semantic information from point clouds enables us to understand a scene, classify objects, and generate high-level models with CAD-like geometries from them. It can also provide a significant improvement to existing algorithms, such as those used to construct Digital Terrain Models (DTMs) from Digital Surface Models (DSMs) (Unger *et al.*, 2009). With the growing popularity of laser scanners, the availability of drones as surveying tools, and the rise of commercial photogrammetry software capable of generating millions of points from images, there exists an increasing need for fully automated extraction of semantic information from this kind of data. Although some of the commercial photogrammetry software available today offer tools such as automated DTM extraction (Pix4Dmapper, 2017, Photoscan, 2017), semantic classification is typically left to specialized software packages (eCognition, 2017, GlobalMapper, 2017) that rely on 2.5D orthomosaics and DSMs as an input.

The need for semantic modeling of 3D point data has inspired many research and application engineers to model specific structures. Often the proposed solutions were handcrafted to the application at hand: buildings have been modeled by using common image processing techniques such as edge detection (Haala *et al.*, 1998, Brenner, 2000) or by fitting planes to point clouds (Rusu *et al.*, 2007); road networks have been modeled by handcrafted features, and DTM algorithms used heuristics about the size of objects to create a DTM from a DSM. While successful and valuable, these approaches are inherently limited since they cannot be easily applied to detect new classes of objects. The huge boost in the performance of machine learning algorithms over the last years allows for more flexible and general learning and classification algorithms. If supervised or semi-supervised learning and especially classification becomes fast and reliable, machine learning approaches to point cloud classification will find their way into common photogrammetric workflows. Therefore, we focus here on machine learning techniques that will allow the users to detect objects categories of their own choice.

In this paper we present a method to classify aerial photogrammetry point clouds. Our approach exploits both geometric and color information to classify individual points as belonging to one of the following classes extracted from the LAS standard: *buildings, terrain, high vegetation, roads, human made objects (HMO)* or *cars*. Unlike previous point cloud classification methods that rely exclusively on geometric features, we show that incorporating color information yields a significant increase in accuracy.

We evaluate our approach on four challenging datasets and show that off-the-shelf machine learning techniques together with our new features result in highly accurate and efficient classifiers that generalize well to unseen data. The datasets used for evaluation are publicly available at *https://pix4d.com/research*.

Moreover, we show that our classification approach can be used to generate accurate Digital Terrain Models, without the need for hand-designed heuristics such as Maximally Stable Extremal Regions (MSER) detection on a Digital Surface Model.

Related Work

Methods used to extract semantic information from point clouds can be split into two groups: those that try to segment coherent objects from a scene, and those that focus on assigning an individual class label to each point. Early works using the first approach often converted the point data into a regular 2.5D height grid so that standard image processing techniques, e.g., edge detection, can be applied (Haala *et al.*, 1998; Haala and Brenner, 1999; Wang and Schenk, 2000). A scan line based approach (Sithole and Vosselman, 2003) was proposed for structure detection in urban scenes. Building extraction approaches typically use geometric primitives during the segmentation step. A multitude of such primitives has been proposed, both in 2D, such as planes and polyhedral (Vosselman *et al.*, 2001; Dorninger and Nothegger, 2007), and in 3D (Lafarge and Mallet, 2012; Xiao and Furukawa, 2014). In Rusu *et al.* (2007) the authors fit sampled parametric models to the data for object recognition. Similarly, Oesau *et al.* (2016) investigates supervised machine learning techniques to represent small indoor datasets with planar models for object recognition.

C. Becker, E. Rosinskaya, E. d'Angelo, and C. Strecha are with Pix4D SA, EPFL Innovation Park, Building F, 1015 Lausanne, Switzerland (carlos.becker@pix4d.com).

N. Häni is with the University of Minnesota.

Photogrammetric Engineering & Remote Sensing
Vol. 84, No. 5, May 2018, pp. 287–295.
0099-1112/18/287–295

doi: 10.14358/PERS.84.5.287

The second type of methods assign a label to each point in the point cloud. Typically this is done with supervised machine learning techniques, requiring training on labeled data from which a classification model is learned and then applied to new, unseen data to predict the label of each point. Binary classification has been explored in environmental monitoring to extract road surfaces (Shu *et al.*, 2016), tree species (Böhm *et al.*, 2016; Liu and Böhm, 2015), land cover (Zhou *et al.*, 2016), and construction sites (Xu *et al.*, 2016).

Several other authors employed a multiclass setting to classify multiple types of objects and structures (Brodu and Lague, 2012; Weinmann *et al.*, 2015a; Hackel *et al.*, 2016), which we adopt in this paper. In particular, we follow the work of (Weinmann *et al.*, 2013), which introduced local geometric features that were used to train a Random Forest (RF) classifier for single terrestrial lidar scans. Their set of features was extended later by Hackel *et al.* (2016). Examples of other feature sets used in the point classification context are *Fast Point Feature Histogram* (FPFH) (Rusu *et al.*, 2009) or *Color Signature of Histogram of Orientations* (SHOT) (Tombari *et al.*, 2010). All these methods use handcrafted features that can be considered suboptimal when compared to more recent deep learning techniques (Hu and Yuan, 2016; Qi *et al.*, 2016), which learn features directly on image or point cloud data. Those approaches have not been considered here, since they require large computational power to train the classifier, and may be restrictive at prediction time, depending on the hardware available.

The ambiguity of the classification task can be minimized by modeling also the spatial correlations between the different class labels. Spatial priors are used in Shapovalov and Velizhev (2011) to classify lidar data and in Niemeyer *et al.*(2014), the authors apply Conditional Random Field (CRF) priors to model different probabilities that neighboring labels can have. While those methods show reasonable classification improvements, they are computationally expensive and not easy to parallelize.

In this paper we extend the work on geometric features by (Weinmann *et al.*, 2013; Hackel *et al.*, 2016) and show that incorporating color information provides a significant boost in prediction accuracy, while keeping a low computational load at prediction time. In the following sections we describe our method and present the results obtained on four photogrammetry datasets.

Method

Our approach combines geometric and color features that are fed to a classifier to predict the class of each point in the point cloud. The geometric features are those introduced in Hackel *et al.*,(2016), which are computed at multiple scales, as soon explained further. To exploit color information, we compute additional color features, based on the color of the respective point and its neighbors.

In the next sections we describe the geometric features introduced in (Hackel *et al.*, 2016). We then show how our color features are computed and discuss implementation details.

Geometric Features

Our approach computes geometric features at different scales to capture details at varying spatial resolutions. Below, we first describe how features are computed for a single scale, and then we show how the scale pyramid is constructed.

We follow the method proposed in Weinmann *et al.*(2013) and later in Hackel *et al.*,(2016). To compute the features for a point x, we first obtain its neighboring points $\mathcal{S}_x = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_k\}$ pkg. This set is used to compute a local 3D structure covariance tensor.

$$C_{\mathbf{x}} = \frac{1}{k}\sum_{i=1}^{k}(\mathbf{p}_i - \hat{\mathbf{p}})(\mathbf{p}_i - \hat{\mathbf{p}})^T \quad (1)$$

Table 1. Our set of geometric (top) and color features (bottom) computed for points in local neighborhood $\mathcal{S}_{\mathbf{x}}$. Geometric features are based on eigenvalues of the local structure tensor, moments around the corresponding eigenvectors, height differences in $\mathcal{S}_{\mathbf{x}}$. Color features include HSV space color of the point of interest and its neighborhood; where $\hat{\mathbf{p}} = \arg\min_{\mathbf{p}} \sum_{i=1}^{k} \| \mathbf{p}_i - \mathbf{p} \|$ the medoid of $\mathcal{S}_{\mathbf{x}}$.

Covariance	Omnivariance	$(\lambda_1 \cdot \lambda_2 \cdot \lambda_3)^{1/3}$
	Eigenentropy	$-\sum_{i=1}^{3} \lambda_i \cdot \ln(\lambda_i)$
	Anisotropy	$(\lambda_1 - \lambda_3)/\lambda_1$
	Planarity	$(\lambda_2 - \lambda_3)/\lambda_1$
	Linearity	$(\lambda_1 - \lambda_2)/\lambda_1$
	Surface variation	λ_3
	Scatter	λ_3/λ_1
	Verticality	$1 - \lvert \langle [0, 0, 1] , \mathbf{e}_3 \rangle \rvert$
Moments	1st order, 1st axis	$\sum_{\mathbf{p}\in\mathcal{S}_{\mathbf{x}}} \langle \mathbf{p} - \hat{\mathbf{p}}, \mathbf{e}_1 \rangle$
	1st order, 2st axis	$\sum_{\mathbf{p}\in\mathcal{S}_{\mathbf{x}}} \langle \mathbf{p} - \hat{\mathbf{p}}, \mathbf{e}_2 \rangle$
	2st order, 1st axis	$\sum_{\mathbf{p}\in\mathcal{S}_{\mathbf{x}}} \langle \mathbf{p} - \hat{\mathbf{p}}, \mathbf{e}_1 \rangle^2$
	2st order, 2st axis	$\sum_{\mathbf{p}\in\mathcal{S}_{\mathbf{x}}} \langle \mathbf{p} - \hat{\mathbf{p}}, \mathbf{e}_2 \rangle^2$
Height	Vertical Range	$z_{\max}\{\mathcal{S}_{\mathbf{x}}\} - z_{\min}\{\mathcal{S}_{\mathbf{x}}\}$
	Height below	$z_{\mathbf{p}} - z_{\min}\{\mathcal{S}_{\mathbf{x}}\}$
	Height above	$z_{\max}\{\mathcal{S}_{\mathbf{x}}\} - z_{\mathbf{p}}$
Color	Point color	$[H_{\mathbf{x}}, S_{\mathbf{x}}, V_{\mathbf{x}}]$
	Neighborhood colors	$\frac{1}{\lvert \mathcal{N}_{\mathbf{x}}(r) \rvert} \sum_{\mathbf{p}\in\mathcal{N}_{\mathbf{x}}(r)} [H, S, V]_{\mathbf{p}}$

The eigenvalues $\lambda 1 \geq \lambda 2 \geq \lambda 3 \geq 0$, unit-sum normalized, and the corresponding eigenvectors $\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$ of $C_{\mathbf{x}}$ are used to compute the local geometry features shown in Table 1.

We have slightly changed the geometry feature set from (Hackel *et al.*, 2016) and removed the sum of eigenvalues because it is constant since the eigenvalues are normalized to unit sum. In addition there are the first and second order moments of the point neighborhood around the eigenvectors which help to identify edges and occlusions. Besides the features based on the eigenvalues and eigenvectors of $C_{\mathbf{x}}$, features based on the z coordinate of the point are used to increase their discriminative power.

Multi-Scale Pyramid

To incorporate information at different scales we follow the multi-scale approach of Hackel *et al.*(2016), which has shown to be more computationally efficient than that of (Weinmann *et al.*, 2015b). Instead of computing the geometric features of Table 1 at a single scale, we successively downsample the original point cloud to create a multi-scale pyramid with decreasing point densities. The geometric features described earlier are computed at each pyramid level and later concatenated.

Pyramid Scale Selection

In order to generalize over different point clouds with varying spatial resolution, we need to choose a fixed set of pyramid levels. This is particularly important when dealing with data with varying Ground Sampling Distance (GSD), which affects the spatial resolution of the point cloud. The GSD is a characteristic of the images used to generate the point cloud. It is the distance between two consecutive pixel centers measured in the orthographic projection of the images onto the Digital Surface Model (DSM). Among other factors, the GSD depends on the altitude from which the aerial photos were taken.

With this in mind, we set the starting resolution of the pyramid to four times the largest GSD in our datasets, or 4 × 5.1cm = 20.4 centimeters. In total we compute 8 scales, with a downsampling factor of 2. With these values we were able to capture changes in patterns of surfaces and objects which vary with distance (e.g., buildings have significant height variations at the scale of dozens meters, while cars, trees do at only a few meters).

Color Features

To increase the discriminative power of the feature set, we combine the geometric features introduced above with color features. Our color features are computed in the HSV color space first introduced by (Smith, 1978), since the analysis of the Pearson product-moment correlation coefficient and the Fisher information of our training data showed that we should expect higher information gain from the HSV over RGB color space.

Besides the HSV color values of the point itself, we compute the average color values of the neighboring points in the original point cloud (i.e., not downsampled). These points are selected as the points within a certain radius around the query point. We experimented with radii of 0.4 m, 0.6 m, and 0.9 m to balance between classification speed and accuracy.

Training and Classification

We use supervised machine learning techniques to train our classifier. We experiment with two well-known ensemble methods: Random Forest (RF) (Breiman, 2001) and Gradient Boosted Trees (GBT) (Friedman *et al.*, 2001). Though RF has been used extensively in point cloud classification (Weinmann *et al.*, 2015b; Hackel *et al.*, 2016), we provide a comparison to GBT and show that the latter can achieve higher accuracies at a similar computational complexity.

Both RF and GBT can generate conditional probabilities and are applicable to multi-class classification problems. They are easily parallelized and are available as reusable software packages in different programming languages.

RF is a very successful learning method that trains an ensemble of decision trees on random subsets of the training data. The output of a RF is the average of the predictions of all the decision trees in the ensemble, which has the effect of reducing the overall variance of the classifier.

On the other hand, the Gradient Boosted Trees (GBT) method trains an ensemble of trees by minimizing its loss over the training data in a greedy fashion (Friedman *et al.*, 2001). GBT has been described as one of the best off-the- shelf classification methods, and it has been shown to perform similarly or better than RF in various classification tasks (Caruana and Niculescu-Mizil, 2006).

Speeding Up Prediction with Early Stopping

To speed up prediction we implement a basic early stopping scheme on top of Gradient Boosted Trees. At prediction time and for each sample, we compute the margin of the most-voted class every e_N evaluated trees. If the margin is larger than a threshold e_τ we assume that the classifier is confident enough about this sample, and therefore there is no need to evaluate the remaining trees in the ensemble. We will show later that this early stopping scheme can speed up prediction by two or three times, improving user experience in interactive applications.

Implementation Details

We implemented our software in C++ to ease its later integration into the Pix4DMapper software. For prototyping and evaluation we used Julia (Bezanson *et al.*, 2014). For fast neighbor search we used the header-only nanoflann library[1] which implements a kd-tree search structure.

The implementation of the RF comes from the ETH Random Forest Library[2]. We parallelized training and prediction, reducing computation times significantly. For GBT we used Microsoft's LightGBM[3] .

Evaluation

In this section we describe first the datasets and classification methods used for the experiments. Next, we show that our implementation is able to reproduce the results presented in (Hackel *et al.*, 2016) on the Paris-rue-Madame dataset. For this dataset we use purely geometric features, as no color information is available. Finally, we evaluate our approach on four challenging aerial photogrammetry point clouds. Our experiments demonstrate that using color information boosts performance significantly, both quantitatively and qualitatively.

Datasets

Table 2 shows the characteristics of the datasets employed for evaluation. The Paris-rue-Madame dataset (Serna *et al.*, 2014) does not contain color information and was solely used to verify that our geometric features perform as well as those of (Hackel *et al.*, 2016)

Table 2. Point cloud datasets used for evaluation.

Dataset	Acquisition	Color	# points
Paris-rue-Madame	Laser Scan	no	20M
Ankeny	Aerial Images	yes	9.0M
Buildings	Aerial Images	yes	3.4M
Cadastre	Aerial Images	yes	5.8M
Rural	Aerial Images	yes	15.4M

Table 3. Point cloud dataset content break down. The datasets are heterogeneous and contain different objects and types of terrain.

Feature	Ankeny	Buildings	Cadastre	Rural
Roads				
Ground/Grass on flatland				
Ground/Grass on slopes	☒	☒		☒
Dry cropland		☒	☒	☒

Our main interest is the aerial photogrammetry and the four last datasets of Table 2. The images were processed with Pix4Dmapper to obtain their respective dense point clouds that were used as the input for our approach. Note that the GSD varies significantly between datasets. A 3D visualization is presented in Figure 3a.

Moreover, each dataset contains different types of objects and terrain surfaces as shown in Table 3. For example, while all datasets contain roads, cropland only appears in one of them. This table will be useful later to analyze the performance of our approach on each dataset.

We have made three photogrammetry datasets publicly available at *https://pix4d.com/research*.

Experimental Setup

To evaluate our method, we test different combinations of feature sets and classifiers on photogrammetry data. We compare two different setups to evaluate the performance of our approach: within the same dataset, or intra-dataset, and across different datasets, or inter-dataset. For training we sampled 50k points of each class at random, resulting in 300k training samples.

1. https://github.com/jlblancoc/nanoflann)
2. http://www.prs.igp.ethz.ch/research/Source code and datasets.html
3. https://github.com/Microsoft/LightGBM

The different feature sets used in our experiments are summarized below:

- Geometric features ($\mathcal{G}$): the geometric eigenvalue-based features shown in Table 1. We use $k = 10$ neighbors to construct $S_{\mathbf{x}}$.
- Geometric features, points color ($\mathcal{C}_p$): HSV color values of the respective 3D point.
- Geometric features, points and neighborhood color ($\mathcal{C}_{\mathcal{N}(r)}$): $\mathcal{C}_p$ set added with averaged HSV color values of the neighboring points within the radius r around the respective 3D point.

Intra-Dataset Experiments

In this setup we divide each dataset into two physically disjoint point clouds. We first find a splitting vertical plane such that the resulting point clouds are as similar as possible with respect to the number of points per class. More specifically, we solve for the vertical plane

$$\hat{p} = \underset{p \in P}{\arg\min} \left[\max_{c \in Y} \left| \frac{1}{\#c} \sum_{\mathbf{x}_i \in p^+} (y_i = c) - \frac{1}{2} \right| \right] \tag{2}$$

where $\#c$ is the number of points of class c in the whole point cloud, Y is the set of all classes present in the point cloud, $p+$ is the set of points falling on one side of the plane p, and P is a set of potential vertical planes of different offsets and rotations.

We then train on one of the splits and test on the other.

Inter-Dataset Experiments

To test the generalization capabilities of our approach to new unseen point clouds we also experiment with a leave-one-out evaluation methodology: we train on two point clouds and test on the remaining one.

Classifier Parameters

For both GBT and RF we used 300 trees, and at each split half of the features were picked at random as possible candidates. For RF the maximum tree depth was set to 30. For GBT we set the maximum number of leaves to 32, learning rate to 0.2, and the bagging fraction to 0.5. These parameters were fixed for all the experiments.

Validation on Laser Scans

In the first experiment we reproduced the results presented on the laser-scan Paris-rue-Madame dataset (Serna *et al.*, 2014). The training and test data sets are generated the same way as in (Weinmann *et al.*, 2015b) and (Hackel *et al.*, 2016) by randomly sampling without replacement 1,000 points per each class for training, and using the rest of the points for testing. When training a RF we achieved overall accuracies of 95.76 percent compared to the reported 95.75 percent in the paper although our per-class results differed slightly. We also observed that this evaluation procedure typically yields overly optimistic accuracies, which are much higher than the expected accuracy on unseen test data. We noticed that such evaluation resembles an inpainting problem: given a few known labeled points in the cloud, estimate the labels of the rest that lie in-between. This gives a bias to the results and does not represent the classifier's ability to generalize to unseen datasets.

To overcome these issues we propose to split the data set into two non-overlapping regions, train on one half and test on the other, as previously described. If the Paris-rue-Madame dataset is split this way our overall accuracy is reduced to 90 percent. We believe this is a less biased estimator of the performance on unseen data, and adopt this strategy to evaluate performance in the rest of our experiments.

It is worth noting that the Paris-rue-Madame dataset contains only small quantities of some classes such as vegetation and human made objects which were found to be harder to classify correctly by (Hackel *et al.*, 2016).

Experiments on Aerial Photogrammetry Data

Intra-Dataset Results

The misclassification errors for different sets of features are presented in Figure 1, where we can see that color features bring a significant improvement. The best results are obtained with the CN (0.6) features. A second important observation is that GBT consistently outperforms RF, in some cases by a large margin.

Inter-Dataset Results

The results are shown in Figure 2. First, there is an overall increase of classification error, in particular for the Cadastre dataset. To analyze the results in more detail, we computed the confusion matrix for the top-three classes that contribute to the misclassification error, as shown in Table 4. We now discuss the result of each dataset in detail.

Ankeny

The classifier performs very well for buildings and roads, as shown in Figure 3b. However it confuses large amounts of ground points as roads. This is not surprising since most of such mistakes occur in croplands, which are not present in any other dataset. Finally, although high vegetation appears in the top-three misclassified classes in Table 4, this is mostly due to

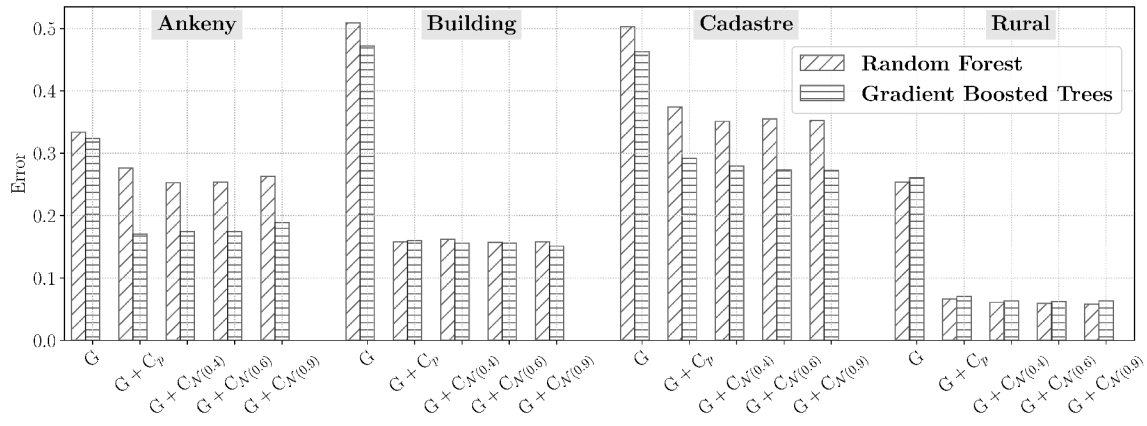


Figure 1. Intra-dataset results: classification error (number of misclassified points divided by overall number of points, the lower the better) when training and testing on different parts of the same dataset. Each dataset is split into physically disjoint training and testing sets through a vertical plane. Incorporating color features CN (0.6) yields a significant improvement. The best results are obtained when combining both geometric and color features.

ambiguities in the ground truth: some bushes were manually labeled as ground, while the classifier predicts them as high vegetation.

Building

The classifier performs very well on this dataset. The highest error is due to predicting buildings as high vegetation or human-made objects. This dataset has the lowest GSD (or highest resolution), and facades of the buildings are well-reconstructed. This is not the case for the other two datasets with higher GSD, where few facade points are available. We hypothesize that the classifier is confused with the facades, finding the vegetation or human-made object to be the closest match.

Cadastre

The classifier predicts vast amounts of ground and vegetation points as buildings and human-made objects, leading to a very high error rate. This result is expected considering Table 3, as the Cadastre dataset contains hills and non-flat ground surfaces, which are not present in any of the other two datasets. Thus, the classifier confuses points in the regions of inclined ground with other classes that are closer in feature space to the training data (e.g. building roofs present a slope that resembles the properties of the points on a hill).

Rural

Our approach performs very well on this dataset, obtaining the lowest misclassification error among all datasets of 6.2 percent. Most of this error is due to vegetation being classified as ground, which is partly because of ambiguities in the ground truth: it is hard to disambiguate high vegetation from ground near the border of a forest.

The analyses above highlight the importance of reliable and varied training data, in that it should resemble the unseen data on which the classifier will be applied, e.g., different landscapes, seasons, shapes of buildings, etc.

Qualitative Results

Figure 3 shows a 3D view of the Ankeny dataset and the respective classified point cloud obtained when using geometric features only, as well as with our approach. Overall the results are very satisfying, especially when one considers the heterogeneity of the different datasets, as discussed earlier.

Timings

Table 5 shows the breakdown of the timings obtained on a 6-core Intel i7 3.4 GHz computer. Our approach is very efficient, taking less than three minutes to classify every point in any of the presented photogrammetry datasets. This makes it ideal for the applications where the user needs to interact

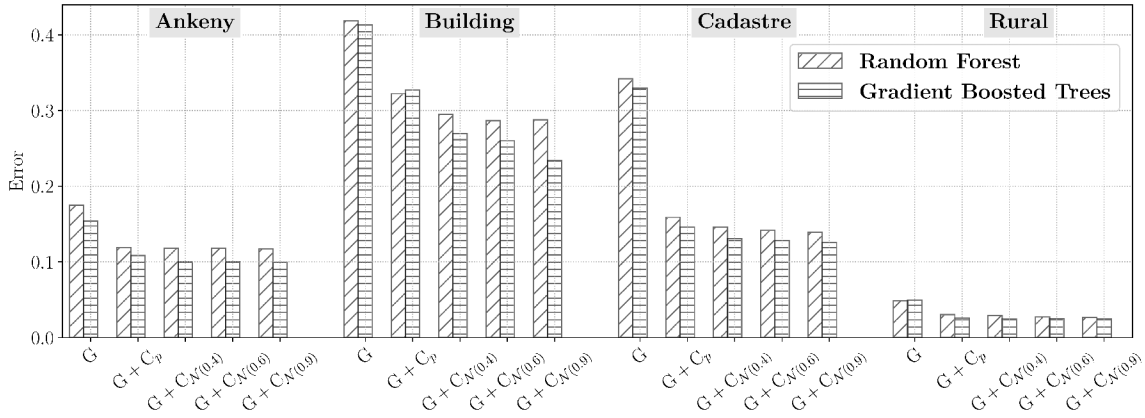
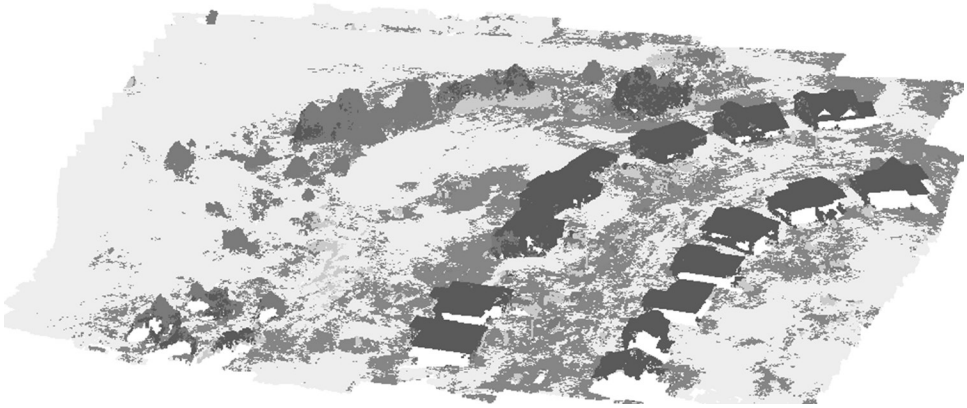


Figure 2. Inter-dataset results: classification error (number of misclassified points divided by overall number of points, the lower the better) when training on two datasets and testing on the remaining one. Similar to Figure 1, the best results are obtained with both geometric and color features CN (0.6). The results on Ankeny and Cadastre show higher classification errors than those obtained when training and testing on the same dataset in Fig. 1. This is due to the lack of training data for certain objects or terrain type.

Table 4. Confusion matrix for the top-three misclassified classes. Results obtained for the G + CN (0.6) features with the GBT classifier, training on two datasets and testing on the remaining third one. Percentages are with respect to the total number of points in the testing dataset.

Dataset			Predicted Label						
Training	**Testing**	**True Label**	**Ground**	**High vegetation**	**Building**	**Road**	**Car**	**HMO**	**Overall error**
Building		Ground	56%	2.3%	0.3%	10%	0.2%	0.6%	12.1%
Cadastre	Ankeny	High vegetation	2.7%	6.7%	0.1%	0.1%	0.1%	0.1%	3.0%
Rural		HMO	0.3%	0.1%	0.1%	0.1%	0.2%	0.3%	0.9%
Ankeny		Road	4.2%	0.1%	0.1%	32%	0.1%	0.4%	4.9%
Cadastre	Building	Building	0.2%	1.8%	30%	0.4%	0.1%	1.7%	4.2%
Rural		High vegetation	2.0%	10%	0.2%	0%	0%	0.2%	2.5%
Ankeny		Ground	36%	4.1%	4.1%	5.1%	1.4%	1.3%	14.8%
Building	Cadastre	Road	3.0%	0.7%	0.4%	15.8%	0.2%	0.9%	5.1%
Rural		High vegetation	0.6%	9.1%	2.1%	0.1%	0.0%	0.8%	3.7%
Ankeny		High vegetation	2.9%	50%	0.1%	0.02%	0%	1.2%	3.2%
Building	Rural	Ground	42%	2.4%	0.02%	0.2%	0%	0%	2.7%
Cadastre		Road	0.2%	0%	0%	1.3%	0%	0%	0.2%

(a) Original data



(b) Classification with geometry features only.

(c) Classification with geometry + color features.

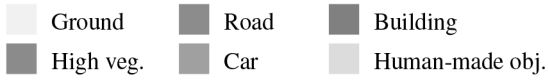


Figure 3. Qualitative results obtained with our approach on the Ankeny dataset, using the other three datasets for training. We used neighbor color features within a 0.6-meter radius neighborhood and the Gradient Boosted Trees classifier. Incorporating color information into the classifier results improves classification, particularly for the roads between buildings.

with the software to correct the training data or fix the classifier's predictions.

Early Stopping

Figure 4 shows the speed ups obtained with early stopping and the reduction in accuracy for all five datasets and different margin thresholds. We set $e_N = 20$ and vary the early stopping threshold e_t. A margin threshold of $e_t = 1.5$ yields a speed up of 2 to 3.5 times compared to not using early stopping, with a very small error increment of 1 percent.

Classification-Based DTM

In this section we show that our classification approach can be used to generate an accurate Digital Terrain Model (DTM). To demonstrate this we employ the variational approach of (Unger *et al.*, 2009), which minimizes an energy functional based on input Digital Surface Model (DSM) and a terrain mask. The goal of this minimization is to smooth and flatten the non-ground areas of the DSM, such as buildings and high vegetation. Ground and non-ground areas are indicated

Table 5. Timings for feature computation, classifier training and prediction. Our whole pipeline runs in less than 3 minutes with any of the provided point clouds, being suitable for interactive applications. This table also shows the benefit of using early stopping with GBT, making prediction between 2 and 4 times faster.

Phase		Ankeny	Building	Cadastre	Rural
Geom. feature extraction		41s	14s	28s	85s
Geom. + Color C_N (0.6) feature extraction		81s	48s	35s	101s
Random Forest (RF)	Train	24m	22m	22m	21m
	Predict	487s	163s	233s	568s
Boosted Trees (GBT)	Train	37s	41s	35s	38s
	Predict (Full)	180s	71s	119s	292s
	Predict (Early Stop, $e_\tau = 1.5$)	73s	23s	48s	78s

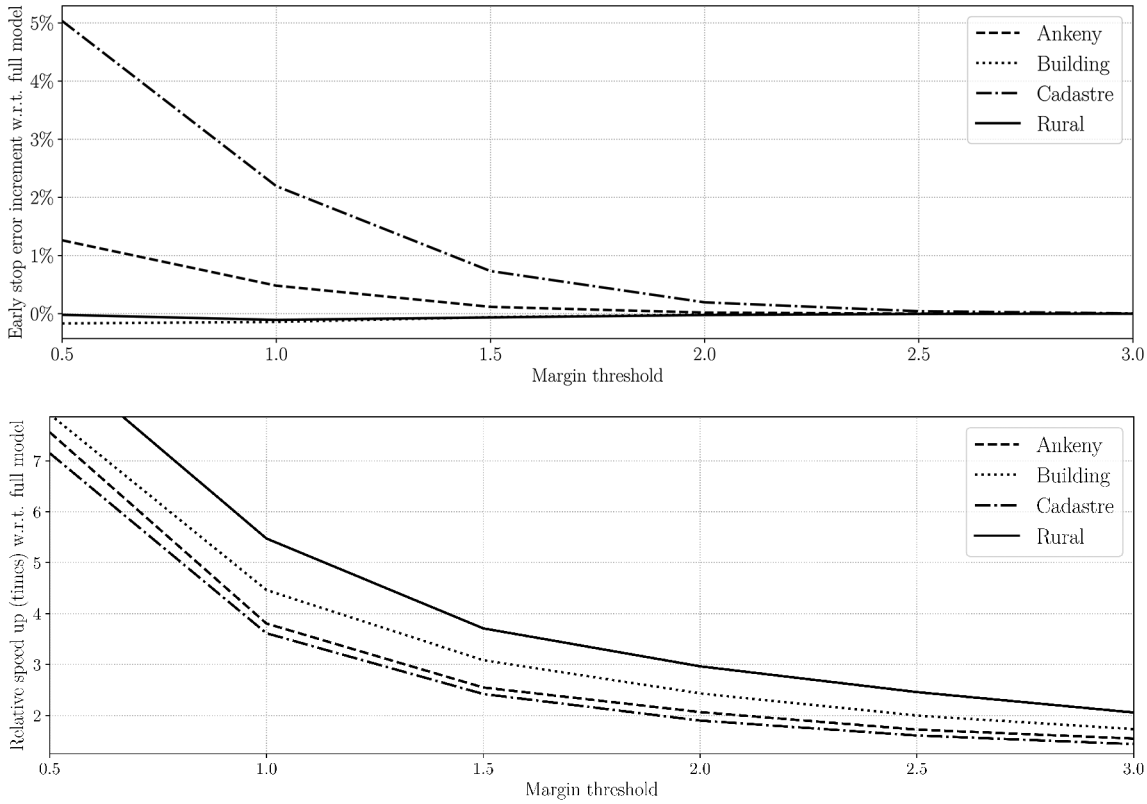


Figure 4. Early stopping: error increment due to early stopping (top) and speed up (bottom). Using a margin threshold of e = 1.5 yields a speed up of 2 to 3.5 times compared to not using early stopping, with less than 1 percent error increment at prediction time.

through a binary terrain mask, which is typically computed using the MSER detector (Matas *et al.*, 2004) on the DSM.

Terrain Mask Generation

We generate the terrain mask in two steps. First, the point cloud is rasterized into an image, and later it is filtered to remove classification artifacts.

Rasterization

Given a classified point cloud, we proceed to discretize the underlying terrain it into a raster image. The resolution of the latter is chosen by the user, according to the desired DTM resolution. For each cell we count the number of points falling into it, and which fraction of that are ground or road surface. If the ratio is greater than 1 we set the respective mask cell to be ground.

Filtering

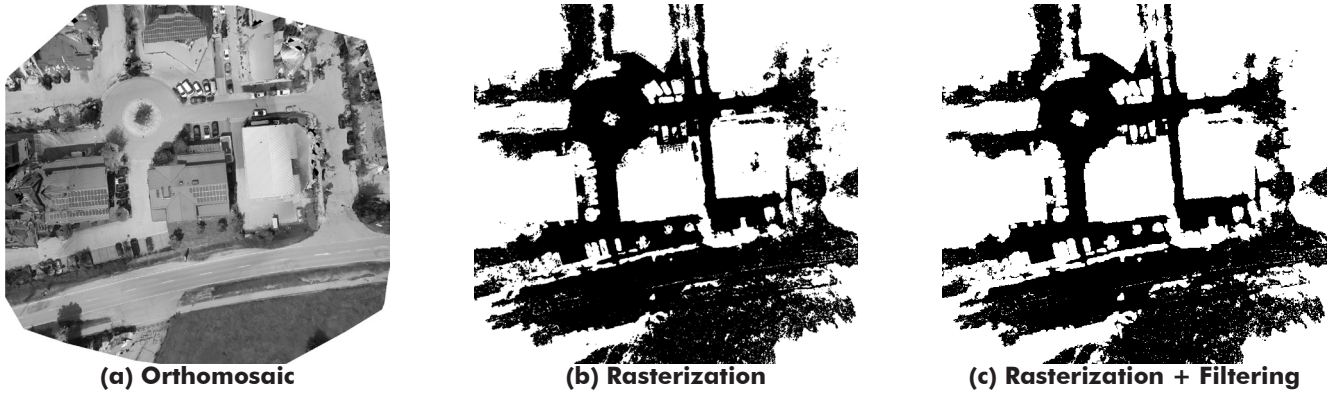
We noticed that sometimes the classifier predicts small patches of road or ground within a roof. We apply the following algorithm to avoid these artifacts from deteriorating the DTM:

1. Generate binary mask as explained above in the rasterization process.
2. Dilate the mask with a 5 × 5 structuring element.
3. Find connected components in the latter.
4. For each connected component c_i .
   a) Evaluate the fraction of pixels $_i$ on the perimeter of ci that are either building or human-made object.
   b) If $\sigma_i > 0.8$ and area(c_i) < 25 m² then remove the pixels inside c_i from the terrain mask.

An example of the results of the rasterization and filtering steps is shown in Figure 5. The filtering step successfully removes patches of ground within roofs and buildings, yielding a more accurate DTM mask.

Evaluation

In this section we compare the DTMs obtained with our approach and with a MSER-based mask. To generate the classification terrain mask we use the models trained during the inter-dataset evaluation scheme in the previous section.



(a) Orthomosaic (b) Rasterization (c) Rasterization + Filtering

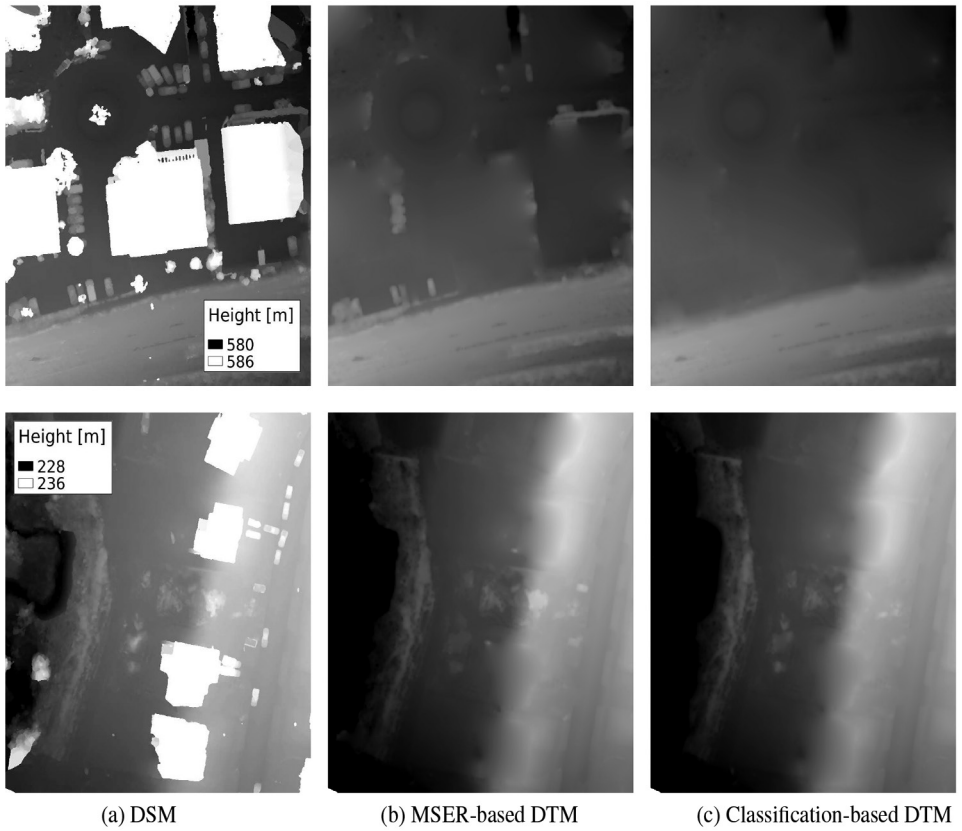
Figure 5. DTM mask generation steps. The rasterized mask in (b) contains a few imperfections that are easily corrected with a basic filtering scheme based on connected component analysis. The final mask shown in (c) no longer contains holes within the buildings.

The results are shown in Figure 6. Although there is no ground truth to perform a quantitative evaluation of the results, the height maps shown therein suggest that the classification-based approach yields more accurate DTMs, as it is able to detect and remove objects such as cars and low vegetation where the MSER technique sometimes fails. This is more evident in the top row of Figure 6b, where some cars and trees were confused as terrain by the MSER method.

Conclusion

In this paper we described an approach for a point-wise semantic labeling of aerial photogrammetry point clouds. The core contribution of our work is the use of color features, what improves significantly the overall classification results. Furthermore, we provide a concise comparison between two standard machine learning techniques that hopefully facilitates the decision making process of future research, showing that the Gradient Boosted Trees classifier outperforms the Random Forest classifier, in some cases by a large margin. Our method performs not only with high accuracies over the whole range of datasets used in the experiments but also with a high computational efficiency, making our approach suitable for interactive applications.

We also showed that our approach can be used to generate accurate Digital Terrain Models, outperforming MSER-based methods and without the need to rely on any additional heuristics. The classification method presented in this paper will soon be part of Pix4Dmapper Pro. Earlier we mentioned that access to properly labeled training data that represents aerial photogrammetry point clouds is limited. To overcome this issue we will implement an incremental training method, where users will be given the possibility to classify their data, visualize and correct errors manually. In a next step we will offer our users the possibility to include their datasets into our training data to improve the classifier quality. As the amount of training data increases we will be able not only to provide more accurate classifiers but to also train specialized ones. For example, we could provide a selection of classifiers for indoor and outdoor scenes, and for different seasons and scales.



(a) DSM (b) MSER-based DTM (c) Classification-based DTM

Figure 6. MSER and classification-based DTM results for the Ankeny and Building datasets. Our classification-base approach removes cars and low vegetation that the MSER-based method is not able to detect, generating a more accurate DTM.

Acknowledgments

This work has been partially supported by European Union's Horizon 2020 DigiArt project No. 665066).

References

Bezanson, J., A. Edelman, S. Karpinski, and V.B. Shah, 2014. Julia: A fresh approach to numerical computing, *arXiv preprint arXiv:1411.1607*.

Böhm, J., M. Bredif, T. Gierlinger, M. Krämmer, R. Lindenberg, K. Liu, F. Michel, and B. Sirmacek, 2016. The IOMULUS urban showcase: Automatic tree classification and identification in huge mobile mapping point points, *ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLI- B3, pp. 301–307.

Breiman, L., 2001. *Random Forests. Machine Learning*, 45(1):5–32.

Brenner, C., 2000. Towards fully automatic generation of city models, *ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pp. 85–92.

Brodu, N., and D. Lague, 2012. 3D terrestrial LIDAR data classification of complex natural scenes using a multi-scale dimensionality criterion: Applications in geomorphology, *ISPRS Journal of Photogrammetry and Remote Sensing*, 68:121–134.

Caruana, R., and A. Niculescu-Mizil, 2006. An empirical comparison of supervised learning algorithms, *Proceedings of the 23rd International Conference on Machine Learning*, ACM, pp. 161–168.

Dorninger, P., and C. Nothegger, 2007. 3D segmentation of unstructured point clouds for building modelling, *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 35(3/W49A):191–196.

eCognition, 2017. Trimble, URL: *www.ecognition.com* (last date accessed: 11 March 2018).

Friedman, J., T. Hastie, and R. Tibshirani, 2001. *The Elements of Statistical Learning*, Vol. 1, Springer series in Statistics, Springer, Berlin.

GlobalMapper, 2017. Blue Marble Geographics, URL: www.bluemarblegeo.com (last date accessed: 11 March 2018.

Haala, N., and C. Brenner, 1999. Extraction of buildings and trees in urban environments, *ISPRS Journal of Photogrammetry and Remote Sensing*, 54, pp. 130–137.

Haala, N., C. Brenner, and K.-H. Anders, 1998. 3D urban GIS from laser altimeter and 2D map data, *International Archives of Photogrammetry and Remote Sensing*, 32, pp. 339–346.

Hackel, T.,J.D. Wegner, and K. Schindler, 2016. Fast semantic segmentation of 3D point clouds with strongly varying density, I*SPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Prague, Czech Republic, 3, pp. 177–184.

Hu, X., and Y. Yuan, 2016. Deep-learning-based classification for DTM extraction from ALS point cloud, *Remote Sensing*, 8(9):730.

Lafarge, F., and C. Mallet, 2012. Creating large-scale city models from 3D point clouds: A robust approach with hybrid representation, *International Journal of Computer Vision*, 99(1):69–85.

Liu, K., and J. Böhm, 2015. Classification of big point cloud data using cloud computing, *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 40(3):553.

Matas, J., O. Chum, M. Urban, and T. Pajdla, 2004. Robust wide-baseline stereo from maximally stable extremal regions, *Image and Vision Computing*.

Niemeyer, J., F. Rottensteiner, and U. Soergel, 2014. Contextual classification of LIDAR data and building object detection in urban areas, *ISPRS Journal of Photogrammetry and Remote Sensing*, 87, pp. 152–165.

Oesau, S., F. Lafarge, and P. Alliez, 2016. Object classification via planar abstraction, *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, III-3, pp. 225–231.

Photoscan, 2017. Agisoft, URL: *www.agisoft.com* (last date accessed: 11 March 2018).

Pix4Dmapper, 2017. Pix4D SA, URL: *www.pix4d.com* (last date accessed: 11 March 2018).

Qi, C.R., H. Su, K. Mo, and L.J. Guibas, 2016. Pointnet: Deep learning on point sets for 3D classification and segmentation, *arXiv preprint arXiv*:1612.00593.

Rusu, R.B., N. Blodow, and M. Beetz, 2009. *Fast Point Feature Histograms (FPFH) for 3D Registration*, IEEE, pp. 3212–3217.

Rusu, R.B., N. Blodow, Z. Marton, A. Soos, and M. Beetz, 2007. Towards 3D object maps for autonomous household robots, *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, pp. 3191–3198.

Serna, A., B. Marcotegui, F. Goulette, and J.-E. Deschaud, 2014. Paris-rue-Madame database: A 3D mobile laser scanner dataset for benchmarking urban detection, segmentation and classification methods, *Proceedings of the 4th International Conference on Pattern Recognition, Applications and Methods - ICPRAM 2014*.

Shapovalov, R., and A. Velizhev, 2011. *Cutting-Plane Training of Non-associative Markov Network for 3D Point Cloud Segmentation*, IEEE, pp. 1–8.

Shu, Z., K. Sun, K. Qiu, and K. Ding, 2016. Pairwise SVM for on-board urban road LIDAR classification, *ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLI-B1, pp. 109–113.

Sithole, G., and G. Vosselman, 2003. Automatic structure detection in a point-cloud of an urban landscape, , 2003. *Proceedings of the 2nd GRSS/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas*, IEEE, pp. 67–71.

Smith, A.R., 1978. Color gamut transform pairs, *ACM SIGGRAPH Computer Graphics*, 12(3):12–19.

Tombari, F., S. Salti, and L. Di Stefano, 2010. Unique signatures of histograms for local surface description, *Proceedings of the European Conference on Computer Vision*, Springer, pp. 356–369.

Unger, M., T. Pock, M. Grabner, A. Klaus, and H. Bischof, 2009. A variational approach to semiautomatic generation of digital terrain models, *Advances in Visual Computing*, pp. 1119–1130.

Vosselman, G., and S. Dijkman, 2001. 3D building model reconstruction from point clouds and ground plans, *International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences*, 34(3/W4):37–44.

Wang, Z., and T. Schenk, 2000. Building extraction and reconstruction from LIDAR data, *International Archives of Photogrammetry and Remote Sensing*, 33(B3/2; PART 3:958–964.

Weinmann, M., B. Jutzi, and C. Mallet, 2013. Feature relevance assessment for the semantic interpretation of 3D point cloud data, *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* II-5/W2:313–318.

Weinmann, M., A. Schmidt, C. Mallet, S. Hinz, F. Rottensteiner, and B. Jutzi, 2015a. Contextual classification of point cloud data by exploiting individual 3D neighborhoods, *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-3/W4:271–278.

Weinmann, M., S. Urban, S. Hinz, B. Jutzi, and C. Mallet, 2015b. Distinctive 2D and 3D features for automated large-scale scene analysis in urban areas, *Computers & Graphics*, 49:47–57.

Xiao, J., and Y. Furukawa, 2014. Reconstructing the worlds museums, I*nternational Journal of Computer Vision*, 110(3);243–258.

Xu, Y., S. Tuttas, L. Heogner, and U. Stilla, 2016. Classification of photogrammetric point clouds of scaffolds for construction site monitoring using Subspace Clustering and PCA, *ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLI-B3:725–732.

Zhou, M., C.R. Li, L. Ma, and H.C. Guan, 2016. Land cover classification from full- waveform LIDAR data based on Support Vector Machines, *ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLI-B3:447–452.

PE&RS

PE&RS 2018 Advertising Rates

PHOTOGRAMMETRIC ENGINEERING & REMOTE SENSING

The official journal for imaging and geospatial information science and technology

Advertise in the #1 publication in the imaging and geospatial information industry!

asprs THE IMAGING & GEOSPATIAL INFORMATION SOCIETY

PE&RS (*Photogrammetric Engineering & Remote Sensing*) is the official journal of the American Society for Photogrammetry and Remote Sensing—the Imaging and Geospatial Information Society (ASPRS). This highly respected publication is the #1 publication in the industry by a wide margin. Advertisers with a desire to reach this market can do no better than *PE&RS*. It delivers the right audience of professionals who are loyal and engaged readers, and who have the purchasing power to do business with you.

***PE&RS* Readership Highlights**

Circulation: 4,000

Total audience: 8,400*

Founded in 1934, the American Society for Photogrammetry and Remote Sensing (ASPRS) is a scientific association serving professional members throughout the world. Our mission is to advance knowledge and improve understanding of mapping sciences to promote the responsible applications of photogrammetry, remote sensing, geographic information systems (GIS), and supporting technologies.

Our members are analysts/specialists, educators, engineers, managers/administrators, manufacturers/ product developers, operators, technicians, trainees, marketers, and scientists/researchers. Employed in the disciplines of the mapping sciences, they work in the fields of Agriculture/Soils, Archeology, Biology, Cartography, Ecology, Environment, Forestry/ Range, Geodesy, Geography, Geology, Hydrology/Water Resources, Land Appraisal/Real Estate, Medicine, Transportation, and Urban Planning/Development.

Reserve your space today!

Bill Spilman, President, Innovative Media Solutions
320 W. Chestnut St., P.O. Box 399, Oneida, IL 61467
(877) 878-3260 toll-free, (309) 483-6467 phone, (309) 483-2371 fax
bill@innovativemediasolutions.com

	Corporate Member Exhibiting at a 2017 ASPRS Conference	Corporate Member	Exhibitor	Non Member
All rates below are for four-color advertisments				
Cover 1	$1,850	$2,000	$2,350	$2,500
In addition to the cover image, the cover sponsor receives a half-page area to include a description of the cover (maximum 500 words). The cover sponsor also has the opportunity to write a highlight article for the journal. Highlight articles are scientific articles designed to appeal to a broad audience and are subject to editorial review before publishing. The cover sponsor fee includes 50 copies of the journal for distribution to their clients. More copies can be ordered at cost.				
Cover 2	$1,500	$1,850	$2,000	$2,350
Cover 3	$1,500	$1,850	$2,000	$2,350
Cover 4	$1,850	$2,000	$2,350	$2,500
Advertorial	1 Complimentary Per Year	1 Complimentary Per Year	n/a	n/a
Full Page	$1,000	$1,175	$2,000	$2,350
2 page spread	$1,500	$1,800	$3,200	$3,600
2/3 Page	$1,100	$1,160	$1,450	$1,450
1/2 Page	$900	$960	$1,200	$1,200
1/3 Page	$800	$800	$1,000	$1,000
1/4 Page	$600	$600	$750	$750
1/6 Page	$400	$400	$500	$500
1/8 Page	$200	$200	$250	$250
Other Advertising Opportunities				
Wednesday Member Newsletter Email Blast	1 Complimentary Per Year	1 Complimentary Per Year	$600	$600

A 15% commission is allowed to recognized advertising agencies

THE MORE YOU ADVERTISE THE MORE YOU SAVE! *PE&RS* offers frequency discounts. Invest in a three-times per year advertising package and receive a 5% discount, six-times per year and receive a 10% discount, 12-times per year and receive a 15% discount off the cost of the package.

Large-Scale Supervised Learning For 3D Point Cloud Labeling: Semantic3d.Net

Timo Hackel, Jan D. Wegner, Nikolay Savinov, Lubor Ladicky, Konrad Schindler, and Marc Pollefeys

Abstract

In this paper we review current state-of-the-art in 3D point cloud classification, present a new 3D point cloud classification benchmark data set of single scans with over four billion manually labeled points, and discuss first available results on the benchmark. Much of the stunning recent progress in 2D image interpretation can be attributed to the availability of large amounts of training data, which have enabled the (supervised) learning of deep neural networks. With the data set presented in this paper, we aim to boost the performance of CNNs also for 3D point cloud labeling. Our hope is that this will lead to a breakthrough of deep learning also for 3D (geo-) data. The semantic3D.net data set consists of dense point clouds acquired with static terrestrial laser scanners. It contains eight semantic classes and covers a wide range of urban outdoor scenes, including churches, streets, railroad tracks, squares, villages, soccer fields, and castles. We describe our labeling interface and show that, compared to those already available to the research community, our data set provides denser and more complete point clouds, with a much higher overall number of labeled points. We further provide descriptions of baseline methods and of the first independent submissions, which are indeed based on CNNs, and already show remarkable improvements over prior art. We hope that semantic3D.net will pave the way for deep learning in 3D point cloud analysis, and for 3D representation learning in general.

Introduction

Neural networks have made a spectacular comeback in image analysis since the seminal paper of (Krizhevsky *et al.,* 2012), which revives earlier work of Fukushima (1980) and LeCun *et al.,* (1989). Especially deep convolutional neural networks (CNNs) have quickly become the core technique for a whole range of learning-based image analysis tasks. The large majority of state-of-the-art methods in computer vision and machine learning now include CNNs as one of their essential components. Their success for image-interpretation tasks is mainly due to (i) easily parallelisable network architectures that facilitate training from millions of images on a single GPU, and (ii) the availability of huge public benchmark data sets like ImageNet (Deng *et al.,* 2009; Russakovsky *et al.,* 2015) and Pascal VOC (Everingham *et al.,* 2010) for RGB images, or SUN RGB-D (Song *et al.,* 2015) for RGB-D data.

While CNNs have been a great success story for image interpretation, they have not yet made a comparable impact for 3D point cloud interpretation. What makes supervised learning hard for 3D point clouds is the sheer size of millions of points per data set, and the irregular, not grid-aligned, and in places very sparse distribution of the data, with strongly varying point density (Figure 1).

While recording point clouds is nowadays straightforward, the main bottleneck is to generate enough manually labeled training data, needed for contemporary (deep)

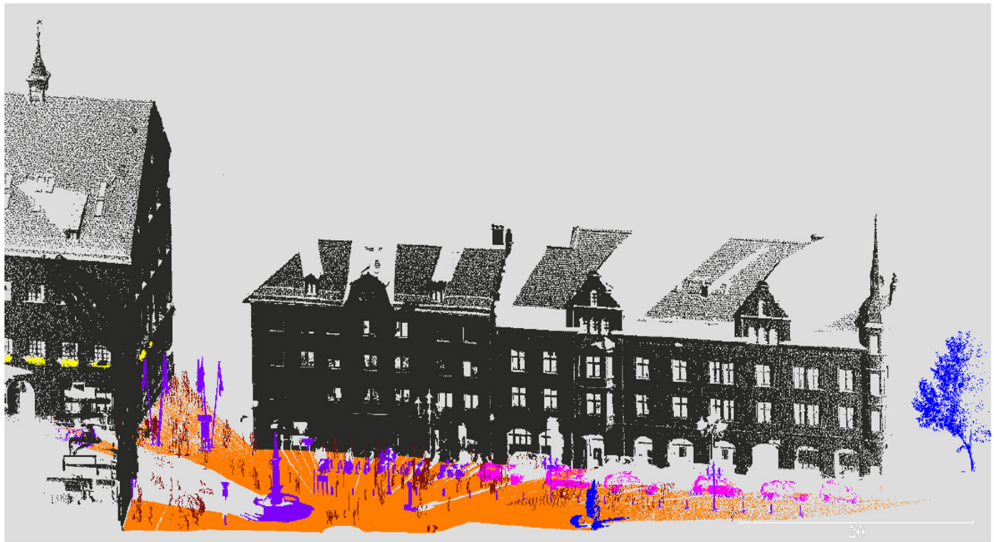


Figure 1. Example point cloud from the benchmark dataset, where colors indicate class labels.

Timo Hackel, Jan D. Wegner, and Konrad Schindler are with IGP, ETH Zurich, Switzerland (jan.wegner@geod.baug.ethz.ch).

Nikolay Savinov, Lubor Ladicky, and Marc Pollefeys are with CVG, ETH Zurich, Switzerland.

Photogrammetric Engineering & Remote Sensing
Vol. 84, No. 5, May 2018, pp. 297–308.
0099-1112/18/297–308

doi: 10.14358/PERS.84.5.297

machine learning to learn good models, that generalize well across new, unseen scenes. Due to the additional dimension, the number of classifier parameters is larger in 3D space than in 2D, and specific 3D effects like occlusion or variations in point density lead to many different patterns for identical output classes. This makes it harder to train good classifiers, so it can be expected that even more training data than in 2D is needed[1]. In contrast to images, which are fairly easy to annotate even for untrained users, 3D point clouds are harder to interpret. Navigation in 3D is more time-consuming and the strongly varying point density aggravates scene interpretation.

In order to accelerate the development of powerful algorithms for point cloud processing[2], we provide the (to our knowledge) hitherto largest collection of individual, non-overlapping terrestrial laser scans with point-level semantic ground truth annotation. In total, it consists of over 4×10^9 points, labeled into 8 classes. The data set is split into training and test sets of approximately equal size, without any overlap between train and test scenes. The scans are challenging, not only due to their realistic size of up to $\approx 4\times10^8$ points per scan, but also because of their high angular resolution and long measurement range, leading to extreme density changes and large occlusions. For convenient use of the benchmark, we provide not only freely available data and ground truth, but also an automated online submission system, as well as evaluation tables for the submitted methods. The benchmark also includes baselines, both for the conventional pipeline consisting of eigenvalue-based feature extraction at multiple scales followed by classification with a random forest, and for a basic deep learning approach. Moreover, we briefly discuss the first submissions to the benchmark, which so far all employ deep learning. This article is an extended version of the conference paper (Hackel *et al.,* 2017). Here, we add a more thorough review of related work, with emphasis on the most recent 3D-CNN methods. We also provide descriptions of the two latest CNN-based submissions, which lead the comparison by a significant margin, and seem to confirm that, also for point cloud analysis, deep learning is the most powerful technology developed to date.

Related Work

Here, we first review traditional methods for point cloud segmentation before discussing novel deep learning-based methods for this task. Finally, we review existing benchmark activities and motivate the introduction of our new 3D point cloud benchmark for semantic segmentation.

Point Cloud Segmentation

Early work on semantic point cloud segmentation transformed the points (recorded from airborne platforms) into other representations such as regular raster height maps, in order to simplify the problem and benefit from the comprehensive toolbox of image processing functions (Hug and Wehr, 1997; Maas, 1999; Haala *et al.*; 1998, Rottensteiner and Briese, 2002; Lodha *et al.*, 2006). Much of the pioneering work on true 3D (i.e., not 2.5D) point cloud processing was developed to guide autonomous outdoor robots (Vandapel *et al.,* 2004; Manduchi *et al.*, 2005; Montemerlo and Thrun, 2006; Lalonde *et al.,* 2006; Munoz *et al.,* 2009b) that rely on laser scanners to acquire data of their surroundings.

In general, it is advantageous if scene interpretation directly operates on 3D points, both for aerial (Charaniya *et al.,* 2004; Chehata *et al.,* 2009; Niemeyer *et al.,* 2011; Yao *et al.,* 2011; Lafarge and Mallet, 2011; Lafarge and Mallet, 2012; Niemeyer *et al.,* 2014; Yan *et al.,* 2015) and for terrestrial data (Brodu and Lague, 2012; Weinmann *et al.,* 2013; Dohan *et al.,* 2015). Full 3D processing can handle data which cannot be reduced to height maps in a straight-forward manner, in particular, terrestrial data generated from multiple scan positions, and mobile mapping data.

Training a good model requires an expressive feature set. A large number of 3D point descriptors has been developed, which typically encode geometric properties within the point's neighborhood, like surface normal orientation, surface curvature, etc. Popular descriptors are for example spin images (Johnson and Hebert, 1999), fast point feature histograms (FPFH) (Rusu *et al.,* 2009) and signatures of histograms (SHOT) (Tombari *et al.,* 2010). One drawback of these rich descriptors is their high computational cost. While computation time is not an issue for small point sets (e.g., sparse key points), it is a crucial bottleneck when all points in a large point cloud shall be classified. A faster alternative, i.e. again for range images rather than true 3D point clouds is the NARF operator, which is popular for key point extraction and description in the robotics community (Steder *et al.,* 2010; Steder *et al.,* 2011). In order to achieve robustness against viewpoint changes, it explicitly models object contour information. A computationally cheaper alternative for full 3D point data are features derived from the 3D structure tensor of a point's neighborhood (Demantke´ *et al.,* 2011), and from the point distribution in oriented (usually vertical) cylinders (Monnier *et al.,* 2012; Weinmann *et al.,* 2013).

Deep Learning for Point Cloud Annotation

Neural networks (usually of the deep, convolutional network flavor) offer the possibility to completely avoid heuristic feature design and feature selection. They are at present immensely popular in 2D image interpretation. Recently, deep learning pipelines have been adapted to voxel grids (Lai *et al.,* 2014; Wu *et al.,* 2015, Maturana and Scherer, 2015) and RGB-D images (Song and Xiao, 2016), also. Being completely data-driven, these techniques have the ability to capture appearance (intensity) patterns as well as geometric object properties. Moreover, their multi-layered, hierarchical architecture has the ability to encode a large amount of contextual information. Deep learning in 3D has been proposed for a variety of applications in robotics, computer graphics, and computer vision. To the best of our knowledge, the earliest attempt that applies a 3D-CNN on a voxel grid is (Prokhorov, 2010). The author classifies objects in lidar point clouds and improves classification accuracy despite limited amount of training data, by combining supervised and unsupervised training. More recent 3D-CNNs that operate on voxel grids include (Maturana and Scherer, 2015) for landing zone detection in 3D lidar point clouds, (Wu *et al.,* 2015) for learning representations of 3D object shapes, and (Huang and You, 2016) to densely label lidar point clouds into seven different object categories. A general drawback when directly applying 3D-CNNs to dense voxel grids derived from originally sparse point clouds is the huge memory overhead for encoding empty space. Computational complexity grows cubically with respect to voxel grid resolution, although high detail would only be needed at object surfaces.

Therefore, more recent 3D-CNNs exploit the sparsity commonly found in voxel grids. One strategy is to resort to an octree representation, where empty space (and potentially also large, geometrically simple object parts) are represented

1. The number of 3D points of semantic3d.net (4×10^9 points) is at the same scale as the number of pixels of the SUN RGB-D benchmark ($\approx 3.3\times10^9$ px) (Song *et al.,* 2015), which aims at 3D object classification. However, the number of 3D points per laser scan ($\approx 4\times10^8$ points), and thus the variability in point density, object scale, etc. is considerably larger than the number of pixels per image ($\approx 4\times10^5$ px).

2. Note that, besides laser scanner point clouds, it is also sometimes preferred to classify point clouds generated using structure-from-motion directly instead of going back to the individual images and then merging the results (Riemenschneider *et al.,* 2014).

at coarser scales than object details (Riegler *et al.,* 2017; Engelcke *et al.,* 2017; Tatarchenko *et al.,* 2017). Since the octree partitioning is a function of the object at hand, an important question is how to automatically adapt to new, previously unseen objects at test time. While (Riegler *et al.*, 2017) assume the octree structure to be known at test time, (Tatarchenko *et al.,* 2017) learn to predict the octree structure together with the labels. This allows generalization to unseen instances of a learned object category, without injecting additional prior knowledge.

Another strategy is to rely only on a small subset of the most discriminative points, while neglecting the large majority of less informative ones (Li *et al.,* 2016; Qi *et al.,* 2017a; Qi *et al.,* 2017b). The idea is that the network learns how to select the most informative points from training data and aggregates information into global descriptors for object shapes using fully-connected layers. This allows for both shape classification and per-point labeling, while using only a small subset of points, resulting in significant speed and memory gains.

Benchmark Initiatives for Point Clouds

Benchmarking efforts have a long tradition in the geospatial data community and particularly in ISPRS. Recent efforts include, for example, the ISPRS-EuroSDR benchmark on High Density Aerial Image Matching[3] that evaluates dense matching methods for oblique aerial images (Haala, 2013, Cavegn *et al.,* 2014) and the ISPRS Benchmark Test on Urban Object Detection and Reconstruction, which contains several different challenges like semantic segmentation of aerial images and 3D object reconstruction (Rottensteiner *et al.,* 2013, Rottensteiner *et al.,* 2014).

In computer vision, very large benchmark datasets with millions of images have become standard for learning-based image interpretation. A variety of datasets have been introduced, many tailored for specific tasks, some serving as basis for annual challenges for several consecutive years (e.g., ImageNet, Pascal VOC). Datasets that aim at boosting research in image classification and object detection heavily rely on images downloaded from the internet. Web-based imagery has been a major driver of benchmarks because no expensive, dedicated photography campaigns have to be accomplished for dataset generation. This makes it possible to scale benchmarks from hundreds to millions of images, although often weakly annotated and with a considerable amount of label noise, that has to be taken into account when working with the data. Additionally, one can assume that internet images constitute a very general collection of images with less bias towards particular sensors, scenes, countries, objects etc. This mitigates overfitting, and enables the training of rich, high-capacity models that nevertheless generalize well.

One of the first successful attempts to object detection in images at very large scale is tinyimages[4] with over 80 million small (32×32 px) images (Torralba *et al.,* 2008). A milestone and still widely used dataset for semantic image segmentation is the famous Pascal VOC[5] dataset and challenge (Everingham *et al.,* 2010), which has been used for training and testing many of the well-known, state-of-the-art algorithms today like (Long *et al.,* 2015; Badrinarayanan *et al.,* 2017). Another, more recent dataset is MSCOCO[6], which contains 300,000 images with annotations that allow for object segmentation, object recognition in context, and image captioning. One of the most popular benchmarks in computer vision today is ImageNet[7] (Deng *et al.,* 2009; Russakovsky *et al.,* 2015), which made Convolutional Neural Networks popular in computer vision (Krizhevsky *et al.,* 2012). It contains > 14×10^6 images organized according to the semantic WordNet hierarchy[8], where words are grouped into sets of cognitive synonyms.

The introduction of the popular, low-cost range sensor Microsoft Kinect gave rise to several large RGB-D image databases. Popular examples are the NYU Depth Dataset V2[9] (Silberman *et al.,* 2012) and SUN RGB-D10 (Song *et al.,* 2015) that provide labeled RGB-D images for object segmentation and scene understanding. Compared to laser scanners, low-cost, structured-light rgb-d sensors have much shorter measurement range, lower resolution, and work poorly outdoors, due to interference of the sunlight with the projected infrared pattern.

To the best of our knowledge, no publicly available dataset with laser scans at the scale of the aforementioned vision benchmarks exists today. Thus, many recent Convolutional Neural Networks that are designed for Voxel Grids (Brock *et al.,* 2016; Wu *et al.,* 2015) resort to artificially generated data from the CAD models of ModelNet (Wu *et al.,* 2015), a rather small, synthetic dataset. As a consequence, recent ensemble methods, e.g., (Brock *et al.,* 2016), reach performance of over 97 percent on ModelNet[10], which clearly indicates that the dataset is either too easy, or too small and already significantly overfitted.

Those few existing laser scan datasets are mostly acquired with mobile mapping devices or robots like DUT1 (Zhuang *et al.,* 2015a), DUT2 (Zhuang *et al.,* 2015b), or KAIST (Choe *et al.,* 2013), which are small (<107 points) and not publicly available. Public laser scan datasets include Oakland (Munoz *et al.,* 2009a) (< 2×10^6 points), the Sydney Urban Objects (De Deuge *et al.,* 2013), Paris-rue-Madame (Serna *et al.,* 2014) and data from the IQmulus & TerraMobilita Contest (Vallet *et al.,* 2015). All have in common that they use 3D lidar data from mobile mapping vehicles, which provides a much lower point density than static scans, like ours. They are also relatively small and localized, and thus prone to overfitting. The majority of today's available point cloud datasets comes without a thorough, transparent evaluation that is publicly available on the internet, continuously updated and that lists all submissions to the benchmark.

With the semantic3D.net benchmark presented in this paper, we attempt to close this gap. It provides a much larger labeled 3D point cloud data set with approximately four billion hand-labeled points, comes with a sound evaluation, and continuously updates submissions. It is the first dataset that allows fully-fledged deep learning on real 3D laser scans, with high-quality, per-point supervision[3].

Data

Our 30 published individual, non-overlapping terrestrial laser scans consist of in total ≈ 4 billion 3D points. Although we would have many more scans that overlap largely with the ones in our benchmark data set and would facilitate co-registration for large scenes, we prefer to keep this for a later extension. The main reason for publishing only individual scans is the huge size per scan (2.72 GB for the largest scan). For the same reason, we did not record multiple echoes per pulse. The data set is split into 15 scans for training that come with labels and 15 scans for testing, where labels are not publicly released and kept by the organizers (see parameters in Tables 1 and 2).

3. http://www.ifp.uni-stuttgart.de/ISPRS-EuroSDR/ImageMatching/index.en.html
4. http://groups.csail.mit.edu/vision/TinyImages/
5. http://host.robots.ox.ac.uk/pascal/VOC/
6. http://mscoco.org/
7. http://www.image-net.org
8. https://wordnet.princeton.edu/
9. http://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html
10. http://rgbd.cs.princeton.edu

Table 1. Parameters of the full resolution semantic-8 training data set. Identical names (left column) with different IDs identify scans of the same scene (but with very low overlap). All ground truth labels together have size 0.01 GB for download. All parameters are also provided on the benchmark website: http://www.semantic3d.net/view_dbase.php?chl=1

Train data set	Number of points	Scene type	Description	Download size [GB]
bildstein1	29 302 501	rural	church in bildstein	0.20
bildstein3	23 765 246	rural	church in bildstein	0.17
bildstein5	24 671 679	rural	church in bildstein	0.18
domfountain1	35 494 386	urban	cathedral in feldkirch	0.28
domfountain2	35 188 343	urban	cathedral in feldkirch	0.25
domfountain3	35 049 972	urban	cathedral in feldkirch	0.23
untermaederbrunnen1	16 658 648	rural	fountain in balgach	0.17
untermaederbrunnen3	19 767 991	rural	fountain in balgach	0.17
neugasse	50 109 087	urban	neugasse in st. gallen	0.32
sg27 1	161 044 280	rural	railroad tracks	1.87
sg27 2	248 351 425	urban	town square	2.72
sg27 4	280 994 028	rural	village	1.59
sg27 5	218 269 204	suburban	crossing	1.25
sg27 9	222 908 898	urban	soccer field	1.22
sg28 4	258 719 795	urban	town square	1.40

Table 2. Parameters of the full resolution semantic-8 testing data set. Identical names (left column) with different IDs identify scans of the same scene (but with very low overlap). All parameters are also provided on the benchmark website: http://www.semantic3d. net/view_dbase.php?chl=1.

Test data set	Number of points	Scene type	Description	Download size [GB]
stgallencathedral1	28 181 979	urban	cathedral in st. gallen	0.22
stgallencathedral3	31 328 976	urban	cathedral in st. gallen	0.22
stgallencathedral6	32 342 450	urban	cathedral in st. gallen	0.22
marketsquarefeldkirch1	23 228 738	urban	market square in feldkirch	0.17
marketsquarefeldkirch4	22 760 334	urban	market square in feldkirch	0.15
marketsquarefeldkirch7	23 264 911	urban	market square in feldkirch	0.15
birdfountain1	36 627 054	urban	fountain in feldkirch	0.25
castleblatten1	152 248 025	rural	castle in blatten	0.24
castleblatten5	195 356 302	rural	castle in blatten	0.70
sg27 3	422 445 052	suburban	houses	2.40
sg27 6	226 790 878	urban	city block	1.27
sg27 8	429 615 314	urban	city center	2.08
sg27 10	285 579 196	urban	town square	1.56
sg28 2	170 158 281	rural	farm	0.94
sg28 5	269 007 810	suburban	buildings	1.35

Submitted results on the test set are evaluated completely automatically on the server and repeated submissions are limited to discourage overfitting on the test set. Train and test data sets are always from different scenes to avoid biasing classifiers and ensure that we verify generalization capability. The data set contains urban and rural scenes, such as, farms, town halls, sport fields, a castle, and market squares. We intentionally selected various different natural and man-made scenes to prevent overfitting of the classifiers. All of the published scenes were captured in Central Europe and depict urban or rural European architecture, as shown in Figure 2. Surveying-grade laser scanners were used for recording these scenes. Colorization was performed in a postprocessing step, by generating high-resolution cubemaps from co-registered camera images. In general, static laser scans have a very high resolution and are able to measure long distances with little noise. Especially compared to point clouds derived using structure-from-motion pipelines or Kinect-like structured light sensors, laser scanners deliver superior geometric data quality.

Scanner positions for data recording were selected as usually done in real field campaigns: only little scan overlap as needed for registration, so that scenes can be recorded in a minimum of time. This free choice of the scanning position implies that no prior assumption based on point density and on class distributions can be made. We publish up to three laser scans per scene that have small overlap. The relative position of laser scans at the same location was estimated from targets.

The choice of output classes in a benchmark, independent of downstream applications, is not obvious. Based on feedback from geospatial industry experts, we use the following eight classes, which are considered useful for a variety of surveying applications: (1) man-made terrain: mostly pavement; (2) natural terrain: mostly grass; (3) high vegetation: trees and large bushes; (4) low vegetation: flowers or small bushes which are smaller than 2 m; (5) buildings: Churches, city halls, stations, tenements, etc.; (6) remaining hardscape: a clutter class with for instance garden walls, fountains, benches, etc.; (7) scanning artifacts: artifacts caused by dynamically moving objects during the recording of the static scan; and (8) cars and trucks. Some of these classes are ill-defined, for instance some scanning artifacts could also go for cars or trucks and it can be hard to differentiate between large and small bushes. Yet, we prefer not to alter the class nomenclature in

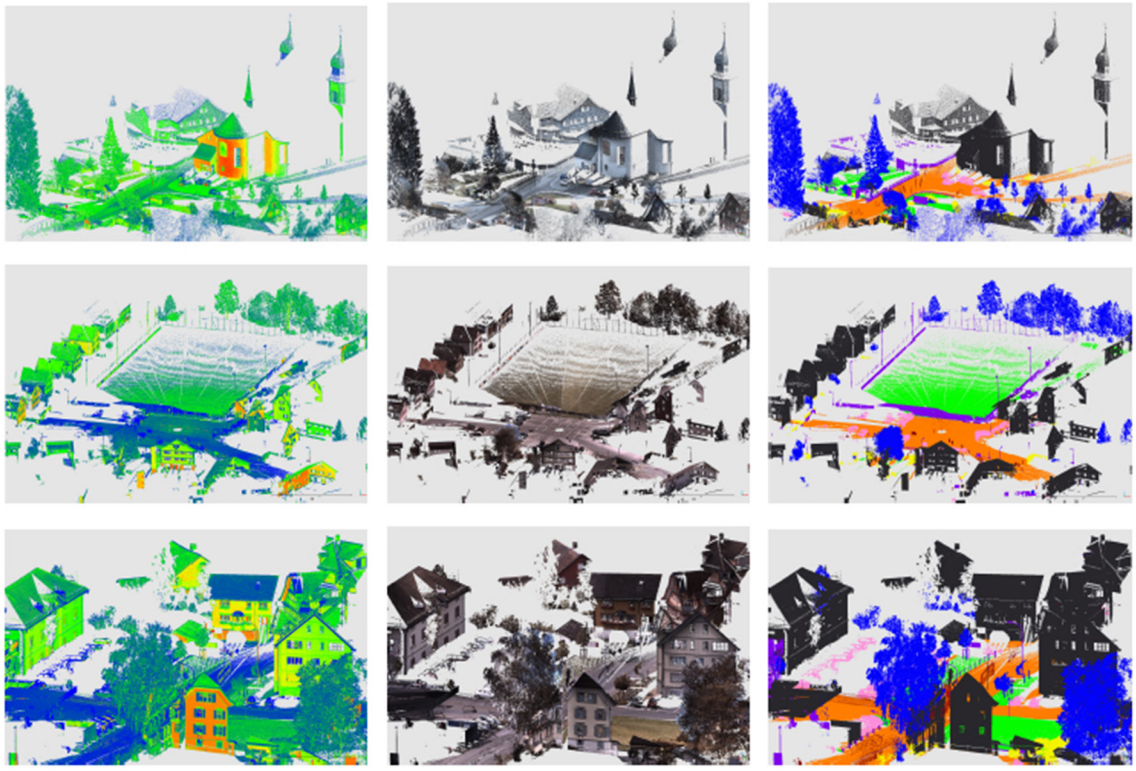


Figure 2. Intensity values (left), RGB colors (middle) and class labels (right) for example data sets.

a way that might reduce ambiguities, but departs from the requirements of the data providers and users. Note also, in many application projects class 7, scanning artifacts, is filtered out in preprocessing with heuristic rule sets. Within the benchmark we prefer to also include that additional classification problem in the overall machine learning pipeline, and thus do not perform any heuristic pre-processing.

In our view, large data sets are important for two reasons: (a) Typically, real world scan data are large. To have an impact on real problems, a method must be able to process large amounts of data, and (b) Large data sets are especially important for modern machine learning methods that involve representation learning (i.e., extracting discriminative low- to high-level features from the raw data). With too small data sets, good results leave strong doubts about possible overfitting; unsatisfactory results, on the other hand, are to interpret as guidelines for further research: are the mistakes due to short-comings of the method, or simply caused by insufficient training data.

Point Cloud Annotation

In contrast to common strategies for 3D data labeling that first compute an automatic over-segmentation and then label segments, we manually assign each point a class label individually. Although this strategy is more labor-intensive, it avoids inheriting errors from the segmentation; and, perhaps more importantly, it ensures that the ground truth does not contain any biases from a particular segmentation algorithm, that could be exploited by the classifier and impair its use with other training data. In general, it is more difficult for humans to label a point cloud by hand than images. The main problem is that it is hard to select a 3D point on a 2D monitor from a set of millions of points without a clear neighborhood/surface structure. We tested two different strategies:

Annotation in 3D

We follow an iterative filtering strategy, where we manually select a couple of points, fit a simple model to the data, remove the model outliers and repeat these steps until all inliers belong to the same class. With this procedure it is possible to select large buildings in a couple of seconds. A small part of the point clouds was labeled with this approach by student assistants at ETH Zurich.

Annotation in 2D

The user rotates a point cloud, fixes a 2D view and draws a closed polygon which splits a point cloud into two parts (inside and outside of the polygon). One part usually contains points from the background and is discarded. This procedure is repeated a few times until all remaining points belong to the same class. In the end, all points are separated into different layers corresponding to classes of interest. This 2D procedure works well with existing software packages (Daniel Girardeau-Montaut, 2016) such that it can be outsourced to external labelers more easily than the 3D work-flow. We used this procedure for all data sets where annotation was outsourced.

Methods

Given a set of points (here: dense scans from a static, terrestrial laser scanner), we want to infer an individual class label per point. We provide three baseline methods that are meant to represent typical categories of approaches recently used for the task, covering the state of the art at the time of creating the benchmark.

2D Image Baseline

We convert color values of the scans to separate images (without depth) with cube mapping (Greene, 1986). Cube maps are centered on the origin of the laser scanner, and thus, we do not experience any self-occlusions. Ground truth labels are

also projected from the point clouds to image space, such that the 3D point labeling task turns into a purely image-based semantic segmentation problem in 2D (Figure 3). We chose the associative hierarchical random fields method (Ladicky *et al.*, 2013) for semantic segmentation because it has proven to deliver good performance for a variety of tasks (e.g., (Montoya *et al.,* 2014; Ladicky´ *et al.,* 2014)) and was available in its original implementation.

The method works as follows: four different types of features – textons (Malik *et al.,* 2001), SIFT (Lowe, 2004), local quantized ternary patters (Hussain and Triggs, 2012), and self-similarity features (Shechtman and Irani, 2007) – are extracted densely at every image pixel. Each feature category is separately clustered into 512 distinct patterns using standard K-means clustering, which corresponds to a typical bag-of-words representation. For each pixel in an image, the feature vector is a concatenation of bag-of-word histograms over a fixed set of 200 rectangles of varying sizes. These rectangles are randomly placed in an extended neighborhood around a pixel. We use multi-class boosting (Torralba *et al.,* 2004) as classifier, and the most discriminative weak features are found as explained in Shotton *et al.* (2006). To add local smoothing without losing sharp object boundaries, the model includes soft constraints that favor constant labels inside superpixels and class transitions at their boundaries. Superpixels are extracted using mean-shift (Comaniciu and Meer, 2002) with three sets of coarse-to-fine parameters as described in (Ladicky *et al.,* 2013). Class likelihoods of overlapping superpixels are predicted using the feature vector consisting of a bag-of-words representation for each superpixel. Pixel-based and superpixel-based classifiers with additional smoothness priors over pixels and superpixels are combined in a conditional random field framework, as proposed in (Kohli *et al.*, 2008). The maximum *a-posteriori* label configuration is found using a graph-cut algorithm (Boykov and Kolmogorov, 2004), with appropriate graph construction for higher-order potentials (Ladicky *et al.*, 2013).

3D Covariance Baseline

The second baseline was inspired by Weinmann *et al.* (2015), and Hackel *et al.* (2016). It infers the class label directly from the 3D point cloud using multiscale features and discriminative learning. Again, we had access to the original implementation of (Hackel *et al.*, 2016). That method uses an efficient approximation of multi-scale neighborhoods, where the point cloud is sub-sampled into a multi- resolution pyramid, such that a constant, small number of neighbors' per level captures the multi-scale information. The multi-scale pyramid is generated by voxel-grid filtering with uniform spacing.

The feature set extracted at each level is an extension of the one described in Weinmann *et al.* (2013). It uses different combinations of eigenvalues and eigenvectors of the covariance per-point neighborhood to represent geometric surface properties. Furthermore, height features based on vertical, cylindrical neighborhoods are added to emphasize the special role of the gravity direction (assuming that scans are, as usual, aligned to the vertical). Note that we do not make use of color values or laser intensities. We empirically found that they did not improve the point cloud classification, moreover color or intensity information is not always available. As classifier, we use a random forest, for which optimal parameters (number of trees and tree depths) are found with grid search and five-fold cross-validation.

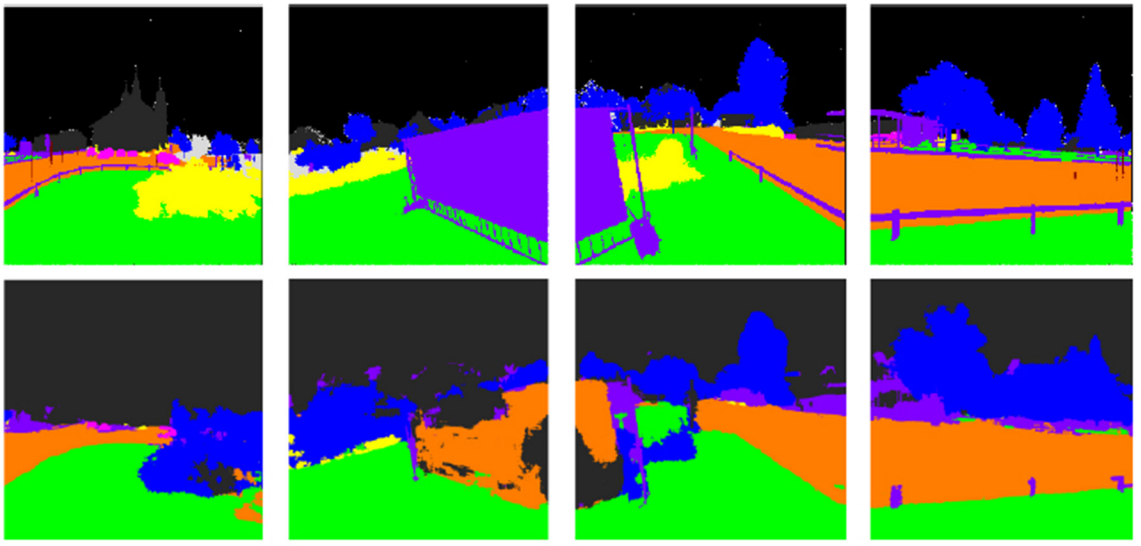


Figure 3. Top row: projection of ground truth to images. Bottom row: results of classification with the image baseline. White: unlabeled pixels, black: pixels with no corresponding 3D point, gray: buildings, orange: man-made ground, green: natural ground, yellow: low vegetation, blue: high vegetation, purple: hard scape, pink: cars.

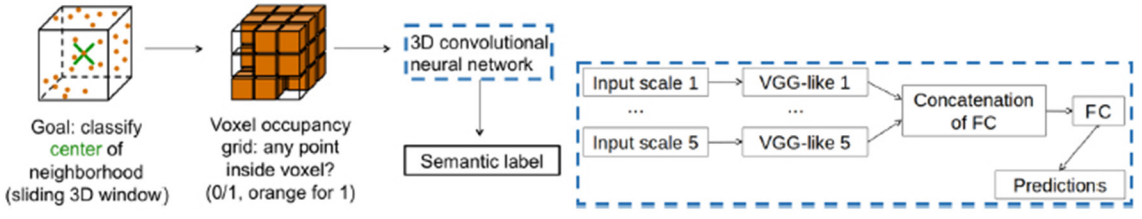


Figure 4. Our deep neural network baseline.

3D CNN Baseline

We design our baseline for the point cloud classification task following recent ideas of VoxNet (Maturana and Scherer, 2015) and ShapeNet (Wu *et al.*, 2015) for 3D encoding. The pipeline is illustrated in Figure 4. Instead of generating a global 3D voxel-grid prior to processing, we create 16 × 16 × 16 voxel cubes per scan point[11]. We do this at 5 different resolutions, with voxel sizes ranging from 2.5 cm to 40 cm (multiplied by powers of 2) and encode empty voxel cells as 0 and filled ones as 1. The input to the CNN is thus encoded in a multidimensional tensor with 5 × 16 × 16 × 16 cube entries per scan point.

Each of the five scales is handled separately by a VGG-like (Simonyan and Zisserman, 2015) network branch that includes convolutional, pooling and ReLU layers. The five separate network paths are finally concatenated into a single representation, which is passed through two fully-connected layers. The output of the second fully-connected layer is an 8-dimensional vector, which contains the class scores for each of the 8 classes in this benchmark challenge. Scores are transformed to class conditional probabilities with the soft-max function.

Before describing the network architecture in detail we introduce the following notation: $c(i, o)$ stands for convolutional layers with 3 × 3 × 3 filters, i input channels, o output channels, zero-padding of size 1 at each border and a stride of 1. $f(i, o)$ stands for fully-connected layers. And, r stands for a ReLU non-linearity, m stands for a volumetric max-pooling with receptive field 2 × 2 × 2, applied with a stride of 2 in each dimension, d stands for a dropout with 0.5 probability, and s stands for a softmax layer.

Our 3D CNN architecture assembles these components to a VGG-like network. We choose the filter size in convolutional layers as small as possible (3 × 3 × 3), as recommended in recent work (He *et al.,* 2016), to have the least amount of parameters per layer and, hence, reduce both the risk of overfitting and the computational cost. Each of the 5 separate network paths, acting at different resolutions, has the sequence:

$$(c(1, 16), \mathbf{r}, m, c(16, 32), \mathbf{r}, m, c(32, 64), \mathbf{r}, m).$$

The output is vectorized, concatenated across all branches (scales), and fed through two fully-connected layers to predict the class responses:

$$(f(2560, 2048), \mathbf{r}, d, f(2048, 8), s).$$

The network is trained by minimizing the standard multi-class cross-entropy loss, with stochastic gradient descent (SGD, (Bottou, 2010)). The SGD algorithm uses randomly sampled mini-batches of several hundred points per batch to iteratively update the parameters of the CNN. We use the popular adadelta (Zeiler, 2012) variant of SGD. We use a mini-batch size of 100 training samples (i.e., points), where each batch is sampled randomly and balanced to contain equal numbers of samples per class. We run training for 74,700 batches and sample training data from a large and representative point cloud with 259 million points (scan sg28 4). A standard preprocessing step for CNNs is data augmentation to enlarge the training set and to avoid overfitting. Here, we augment the training set with a random rotation around the z-axis after every 100 batches. During experiments it turned out that additional training data did not improve performance. This indicates that in our case we rather face underfitting (as opposed to overfitting), i.e., our model lacks the capacity to fully capture all the evidence in the available training data[12]. We thus refrain from further possible augmentations like randomly missing points or adding noise. The network is implemented in C++ and Lua and uses the Torch7 framework (Collobert *et al.*, 2011) for deep learning. Code and documentation are available at *https://github.com/nsavinov/semantic3dnet*.

Submissions to the Benchmark

The two top-performing approaches (Boulch *et al.,* 2017, Lawin *et al.,* 2017) submitted to the benchmark so far[13] both project 3D point clouds to 2D images, so as to harness the strength of well-established CNN models in 2D space. Their strategy is to: (i) render virtual 2D images from viewpoints in the 3D point cloud; (ii) perform semantic classification on the 2D images; (iii) lift the results back into 3D space, and merge the predictions from different 2D views. In the following, we provide a brief overview of both methods. Schematic work-flows are shown in Figures 5 and 6. The currently top-performing method is SnapNet (Boulch *et al.,* 2017). The processing pipeline consists of four main parts (Figure 5):

1. Point clouds are down-sampled with a voxel grid filter, 3D features are extracted (e.g., the deviation of surface normals to a vertical vector, sphericity, etc.), and 3D meshes are generated by running the surface reconstruction approach of Marton *et al. (*2009).
2. Virtual images are rendered from meshes at a high number (400 per point cloud for training) of different camera positions. RGB images as well as composite images with a channel for depth, the deviation of surface normals and

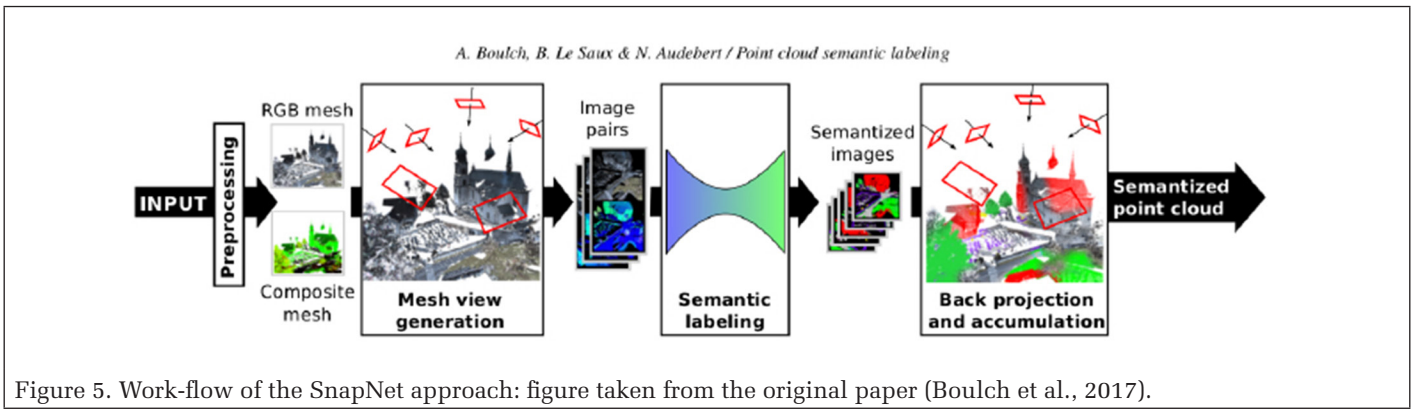


Figure 5. Work-flow of the SnapNet approach: figure taken from the original paper (Boulch et al., 2017).

11. This strategy automatically centers each voxel-cube per scan point. Note that for the alternative approach of a global voxel grid, several scan points could fall into the same grid cell in dense regions of the scan. This would require scan point selection per grid cell, which is computationally costly and results in (undesired) down-sampling.
12. Our model reaches the hardware limits of our GPU (TitanX with 12GB of RAM), we thus did not experiment with larger networks at this point.
13. As of 28 August 2017.

sphericity are computed. For training and validation sets also virtual ground truth images are rendered. The authors propose to select camera view points either randomly in the bounding box of the scene (altitudes vary between 10 and 30 meters above ground) or to apply a multi-scale strategy, where three camera poses are generated for a subset of points that vary in distance to the selected point. A 3D mesh viewer renders virtual 2D images from the mesh.

3. Two different encoder-decoder CNNs, SegNet (Badrinarayanan *et al.,* 2017) and U-Net (Ronneberger *et al.,* 2015), are compared for semantic labeling of the rendered virtual images. Moreover, different strategies to combine RGB and depth information are tested, for example, model averaging and adding a shallow network to the output of the two separate depth and RGB networks.
4. Class responses of the neural network are back-projected to the mesh and averaged over the different virtual views. Finally, a kd-tree is used to assign the class label with the highest class response in the mesh to close points in the point cloud. The overall best results (i.e., those reported for the benchmark, cf. Tables 3 and 4) are obtained with a combination of U-Net, shallow network for depth and RGB fusion, and multi-scale view generation.

The second-best submission at present is DeePr3SS (Lawin *et al.,* 2017) (Figure 6), which follows a conceptually similar strategy as (Boulch *et al.,* 2017):

1. Virtual images with RGB channels as well as channels for depth and surface normals are rendered directly from the point clouds by point splatting (Zwicker *et al.,* 2001) (which, unlike (Boulch *et al.,* 2017), works without an intermediate mesh generation step). In total, 120 camera views are rendered per point cloud by rotating the camera around four vertical axis in the scene. Low quality images are discarded by using two filter strategies: First, images with a coverage below a threshold are removed. Second, views which are too close to large objects are neglected by thresholding the percentage of small depths.
2. Semantic segmentation is performed using fully convolutional networks, where the different inputs are fused by using a multi-stream architecture (Simonyan and Zisserman, 2014) that averages the output of the different streams. The authors use pre-trained VGG16 networks (Simonyan and Zisserman, 2015) for each stream and experiment with different combinations of streams for RGB, depth and normal channels. As often done, pre-training is performed on the ImageNet dataset (Russakovsky *et al.,* 2015).
3. Finally, class responses of the CNN are back-projected to the point cloud. Mapping between 3D points and pixels in the virtual images is given by rendering with point splatting. Class responses of the CNN for all pixels which correspond to the same 3D point are summed up, and the maximum average class response is used as final class label. The authors report that the multi-stream architecture with streams for all RGB, depth and normal channels works best (that workflow is used to produce numbers shown in Table 4 for the benchmark) for DeePr3SS.

Evaluation

We follow the Pascal VOC challenge (Everingham *et al.,* 2010) and choose the Intersection over Union (IoU), averaged over all classes, as our principal evaluation metric.14 . Let the classes be indexed with integers from {1, …, N}, with N the number of different classes. Let C be an N × N confusion matrix of the chosen classification method, where each entry cij is a number of samples from ground-truth class i predicted as class j. Then the evaluation measure per class i is defined as:

$$IoU_i = \frac{c_{ii}}{c_{ii} + \sum_{j \neq 1} c_{ij} + \sum_{k \neq 1} c_{ki}} \quad (1)$$

The main evaluation measure of our benchmark is thus N.

$$IoU = \frac{1}{N} \sum_{i=1}^{n} IoU_i \quad (2)$$

IoU compensates for different class frequencies as opposed to, for example, overall accuracy that does not balance different class frequencies, thus giving higher influence to large classes.

We also report IoU_i for each class i and overall accuracy:

Table 3. Semantic3D benchmark results on the full data set: 3D covariance baseline TMLC-MS, 2D RGB image baseline TML-PC, and first submissions HarrisNet and SnapNet. IoU for categories (1) man-made terrain, (2) natural terrain, (3) high vegetation, (4) low vegetation, (5) buildings, (6) hard scape, (7) scanning artifacts, (8) cars. * Scanning artifacts were ignored for 2D classification because they are not present in the image data.

Method	$\overline{IoU}$	OA	t[s]	IoU_1	IoU_2	IoU_3	IoU_4	IoU_5	IoU_6	IoU_7	IoU_8
SnapNet	0.674	0.910	unknown	0.896	0.795	0.748	0.561	0.909	0.365	0.343	0.772
HarrisNet	0.623	0.881	unknown	0.818	0.737	0.742	0.625	0.927	0.283	0.178	0.671
TMLC-MS	0.494	0.850	38421	0.911	0.695	0.328	0.216	0.876	0.259	0.113	0.553
TML-PC	0.391	0.745	unknown	0.804	0.661	0.423	0.412	0.647	0.124	0.0*	0.058

Table 4. Semantic3D benchmark results on the reduced data set: 3D covariance baseline TMLC-MSR, 2D RGB image baseline TML-PCR, and our 3D CNN baseline DeepNet. TMLC-MSR is the same method as TMLC-MS, the same goes for TMLC-PCR and TMLC-PC. In both cases R indicates classifiers on the reduced dataset. IoU for categories (1) man-made terrain, (2) natural terrain, (3) high vegetation, (4) low vegetation, (5) buildings, (6) hard scape, (7) scanning artifacts, (8) cars. * Scanning artifacts were ignored for 2D classification because they are not present in the image data.

Method	$\overline{IoU}$	OA	t[s]	IoU_1	IoU_2	IoU_3	IoU_4	IoU_5	IoU_6	IoU_7	IoU_8
SnapNet	0.591	0.886	3600	0.820	0.773	0.797	0.229	0.911	0.184	0.373	0.644
DeePr3SS	0.585	0.889	Un-known	0.856	0.832	0.742	0.324	0.897	0.185	0.251	0.592
TMLC-MSR	0.542	0.862	1800	0.898	0.745	0.537	0.268	0.888	0.189	0.364	0.447
DeepNet	0.437	0.772	64800	0.838	0.385	0.548	0.085	0.841	0.151	0.223	0.423
TML-PCR	0.384	0.740	unknown	0.726	0.73	0.485	0.224	0.707	0.050	0.0*	0.15

$$OA = \frac{\sum_{i=1}^{N} c_{ii}}{\sum_{j=1}^{N}\sum_{k=1}^{N} c_{jk}} \quad (3)$$

as auxiliary measures and provide the confusion matrix C . Finally, each participant is asked to specify the time T it took to classify the test set as well as the hardware used for experiments. The computation time (if available) is important to understand how suitable the method is in real-world scenarios, where usually billions of points are required to be processed.

For computationally demanding methods we additionally provide a reduced challenge, consisting of a subset of the original test data. The results of our baseline methods as well as submissions are shown in Table 3 for the full challenge and in Table 4 for the reduced challenge. Of the three published baseline methods the classical machine learning pipeline with hand-designed, covariance-based features performs better than simplistic color image labeling without 3D information, and it also beats our simple CNN baseline, DeepNet. Due to its computational cost we could only run the DeepNet on the reduced data set. We note that DeepNet is meant as a baseline for "naive" application of CNNs to point cloud data, we do expect a more sophisticated, higher-capacity network to perform significantly better. Both SnapNet and DeePr3SS comfortably beat all baselines.

On the full challenge, two CNN methods, SnapNet and HarrisNet (unfortunately unpublished), already beat our best baseline by a significant margin (Table 3) of 12 respective 18 percent points. This indicates that deep learning seems to be the way to go also for point clouds, if enough training data is available. However, it should be noted that both SnapNet and HarrisNet are no true 3D-CNN approaches in the sense that they do not process 3D data directly. Both methods side-step 3D processing and cast semantic segmentation of point clouds as a 2D image labeling problem. For the future of the benchmark, it will be interesting how true 3D-CNN approaches like Riegler *et al. (*2017), Tatarchenko *et al.* (2017), Qi *et al.*(2017a) will perform. As a lesson learned, a future update of the benchmark should include multi-station point clouds that challenge the reprojection strategy.

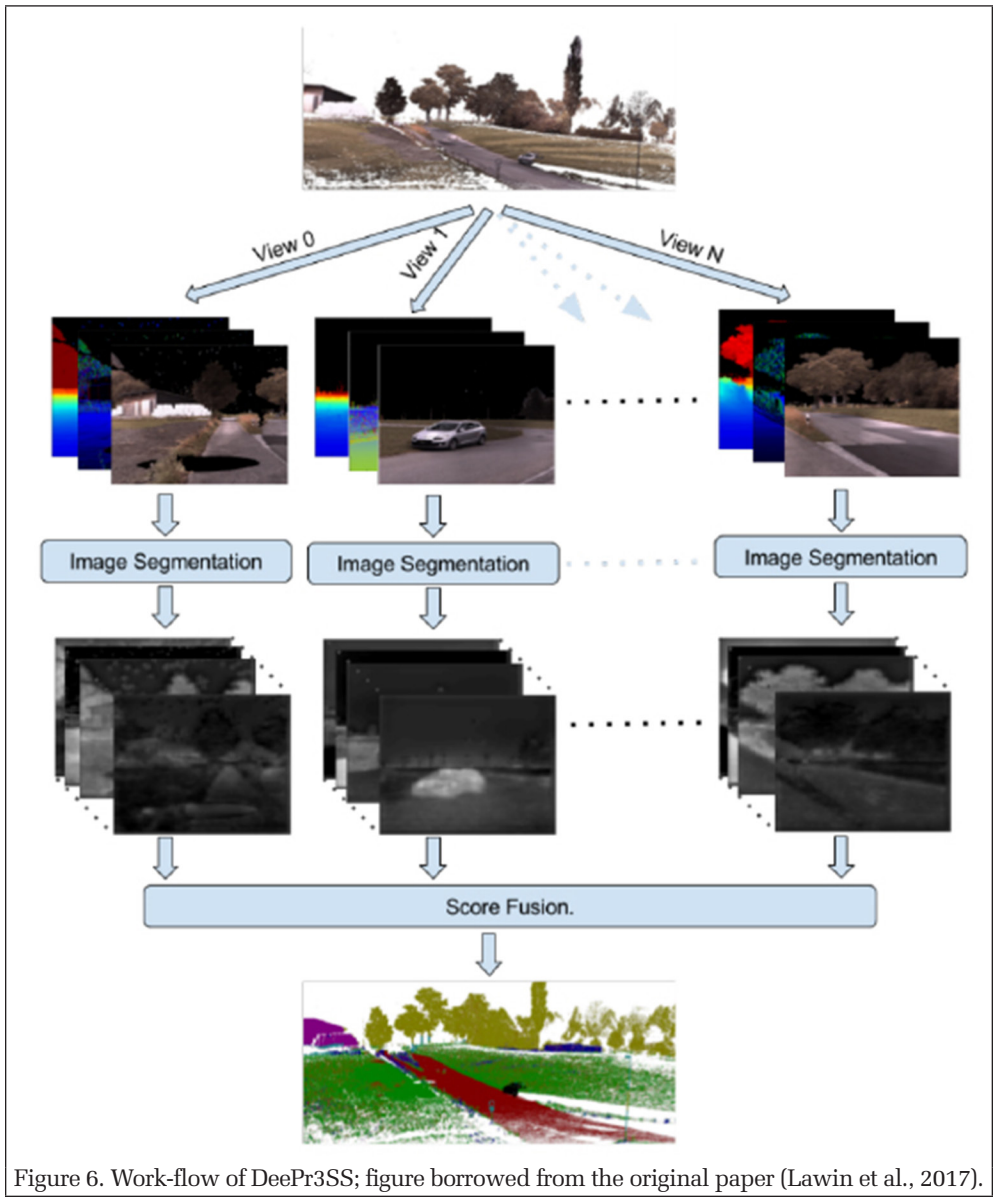


Figure 6. Work-flow of DeePr3SS; figure borrowed from the original paper (Lawin et al., 2017).

Benchmark Statistics

Class distributions in the test and training sets are rather similar, as shown in Figure 7a. Interestingly, the class with most samples is man-made terrain because, out of convenience, operators in the field tend to place the scanner on flat and paved ground. Recall also the quadratic decrease of point density with distance to the scanner, such that many samples are close to the scanner. The largest difference between samples in test and training sets occurs for class building. However, this does not seem to affect the performance of the submissions so far. The most difficult classes, scanning artifacts and cars, have only few training and test samples and a large variation of possible object shapes. Scanning artifacts is probably the hardest class because the shape of artifacts mostly depends on the movement of objects during the scanning process. Note that, following discussions with industry professionals, the class hard scape was designed as a sort of "clutter class" that contains all sorts of man-made objects except for buildings, cars and the ground.

In order to quantify the quality of the manually acquired labels, we also checked the label agreement among human annotators. This check provides an indicative measure how well different annotators agree on the correct labeling, and can be viewed as an internal check of manual labeling precision. To estimate the label agreement between different human annotators, we inspect areas where different scans of the same scene overlap (recall that

overlaps of adjacent scans can be established precisely, using artificial markers placed in the scenes). Since we cannot rule out that some overlapping area might have been labeled twice by the same person (labeling was outsourced, and we thus do not know exactly who annotated what), the observed consistency might in the worst case be slightly too optimistic. Even if scan alignments would be perfect without any error, no exact point-to-point correspondences exist between two scans, because scan points acquired from two different locations will not fall exactly onto the same 3D location.

We thus have to resort to nearest-neighbor search to find point correspondences. Moreover, not all scan points have a corresponding point in the adjacent scan. A threshold of 5 cm on the distance is used to ignore those points where no correspondence exists. Once point correspondences have been established, it is possible to transfer the annotated labels from one point cloud to the other and compute a confusion matrix. Note that this definition of correspondence is not symmetric, "forward" point correspondences from cloud A to cloud B are not in all cases the same as "backward" correspondences from cloud B to cloud A. For each pair, we calculate two intersection-over-union (IoU_i) values, which indicate negligible differences between forward and backward matching, an overall disagreement <3%, and a maximum label disagreement for the worst class (low vegetation) of <5%; see Figure 7b. Obviously, no correspondences between asynchronously acquired scans can be found on moving objects, so we ignored the class scanning artifacts in the evaluation.

Conclusions and Outlook

The *semantic3D.net* benchmark provides a large set of high quality, individual terrestrial laser scans with over 4 billion manually annotated points and a standardized evaluation framework. The data set has been published recently and the first results have been submitted. These already show that deep learning, and in particular, appropriately adapted and well-engineered CNNs, outperform the leading conventional approaches, such as our covariance baseline, on large 3D laser scans. Interestingly, both top-performing methods SnapNet and HarrisNet are no true 3D-CNN approaches in the sense that they do not process 3D data directly. Both methods cast semantic point cloud segmentation as a 2D image labeling problem. This leaves room for methods that directly work in 3D, and we hope to see more submissions of this kind in the future. We are confident that, as more submissions appear, the benchmark will enable objective comparisons and yield new insights into strengths and weaknesses of different classification approaches for point clouds, and that the common test bed can help to guide future research efforts. We hope that the benchmark meets the needs of the research community and becomes a central resource for the development of new, more efficient and more accurate methods for semantic data interpretation in 3D space.

Acknowledgments

This work is partially funded by the Swiss NSF project 163910, the Max Planck CLS Fellowship and the Swiss CTI project 17136.1 PFES-ES.

References

Badrinarayanan, V., A. Kendall, and R. Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 39(12):2481–2495.

Bottou, L., 2010. Large-scale machine learning with stochastic gradient descent, *Proceedings of COMPSTAT'2010*, Springer, pp. 177–186.

Boulch, A., B. Le Saux, and N. Audebert, 2017. Unstructured point cloud semantic labeling using deep segmentation networks, *Proceedings of the Eurographics Workshop on 3D Object Retrieval*, The Eurographics Association, pp. 770–778.

Boykov, Y., and V. Kolmogorov, 2004. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1124–1137.

Brock, A., T. Lim, J. Ritchie, and N. Weston, 2016. Generative and discriminative voxel modeling with convolutional neural networks, *Proceedings of the 3D Deep Learning Workshop at NIPS*.

Brodu, N., and D. Lague, 2012. 3D terrestrial lidar data classification of complex natural scenes using a multi-scale dimensionality criterion: Applications in geomorphology, *ISPRS Journal of Photogrammetry and Remote Sensing*, 68, pp. 121–134.

Cavegn, S., N. Haala, S. Nebiker, M. Rothermel, and P. Tutzauer, 2014. Benchmarking high density image matching for oblique airborne imagery, *ISPRS Archives of Photogrammetry, Remote Sensing, and Spatial Information Sciences*, Vol. XL-3, pp. 45– 52.

Charaniya, A.P., R. Manduchi, and S.K. Lodha, 2004. Supervised parametric classification of aerial lidar data, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*.

Chehata, N., L. Guo, and C. Mallet, 2009. Airborne lidar feature selection for urban classification using random forests. *ISPRS Archives of Photogrammetry, Remote Sensing, and Spatial Information Sciences*, Vol. 38, Part 3/W8, pp. 207–212.

Choe, Y., I. Shim, and M.J. Chung, 2013. Urban structure classification using the 3D normal distribution transform for practical robot applications, *Advanced Robotics*, 27(5), pp. 351–371.

Collobert, R., K. Kavukcuoglu, and C. Farabet, 2011. Torch7: A Matlab-like environment for machine learning, *Proceedings of the Big Learn NIPS Workshop*.

Comaniciu, D., and P. Meer, 2002. Mean shift: A robust approach toward feature space analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5), pp. 603–619.

Daniel Girardeau-Montaut, 2016. The CloudCompare Project,URL: *http://www.danielgm.net/cc/* (last date accessed: 26 March 2018).

De Deuge, M., A. Quadros, C. Hung, and B. Douillard, 2013. Unsupervised feature learning fo classification of outdoor 3D scans, *Proceedings of the Australasian Conference on Robitics and Automation*, Vol. 2.

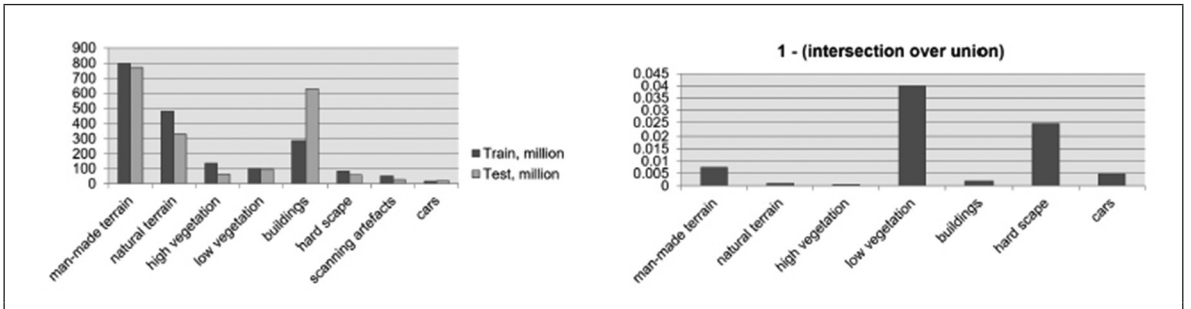


Figure 7. (a) Number of points per class over all scans, and (b) ground truth label errors estimated in overlapping parts of adjacent scans.

Demantké, J., C. Mallet, N. David, and B. Vallet, B., 2011. Dimensionality based scale selection in 3D lidar point clouds, *ISPRS Archives of Photogrammetry, Remote Sensing, and Spatial Information Sciences*, Vol. 38, Part 5/W12, pp. 97–102.

Deng, J., W. Dong, R. Socher, L,-J Li, K. Li, and L. Fei-Fei, 2009. Imagenet: A large-scale hierarchical image database, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255.

Dohan, D., B. Matejek, and T. Funkhouser, 2015. Learning hierarchical semantic segmentations of lidar data, *Proceedings of the International Conference on 3D Vision*, pp. 273–281.

Engelcke, M., D. Rao, D.Z. Wang, C.H. Tong, and I. Posner, 2017. Vote3deep: Fast object detection in 3D point clouds using efficient convolutional neural networks, *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 1355–1361. Everingham, M., L. van Gool, C. Williams, J. Winn, and A. Zisserman, 2010. The Pascal visual object classes (voc) challenge, *International Journal of Computer Vision*, 88(2), pp. 303–338.

Fukushima, K., 1980. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, *Biological Cybernetics*, 36(4), pp. 193–202.

Greene, N., 1986. Environment mapping and other applications of world projections, *IEEE Computer Graphics and Applications*, 6(11), pp. 21–29.

Haala, N., 2013. The landscape of dense image matching algorithms, *Proceedings of Photogrammetric Week 13*, pp. 271–284.

Haala, N., C. Brenner, and K.-H Anders, K, 1998. 3D urban GIS from laser altimeter and 2D map data. *ISPRS Archives of Photogrammetry, Remote Sensing, and Spatial Information Sciences*, 32, pp. 339–346.

Hackel, T., N. Savinov, L. Ladicky, J.D. Wegner, K. Schindler, and M. Pollefeys, 2017. SEMANTIC3D.NET: A new large-scale point cloud classification benchmark, *ISPRS Archives of Photogrammetry, Remote Sensing, and Spatial Information Sciences*, Vol. IV-1-W1, pp. 91–98.

Hackel, T., J.D. Wegner, and K. Schindler, 2016. Fast semantic segmentation of 3D point clouds with strongly varying point density, *ISPRS Archives of Photogrammetry, Remote Sensing, and Spatial Information Sciences*, Vol. III-3, pp. 177–184.

He, K., X. Zhang, S. Ren, and J. Sun, 2016. Deep residual learning for image recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778.

Huang, J., and S. You, 2016. Point cloud labeling using 3D convolutional neural network, *Proceedings of the International Conference on Pattern Recognition*, pp. 2670–2675.

Hug, C., and A. Wehr, 1997. Detecting and identifying topographic objects in imaging laser altimeter data, *ISPRS Archives of Photogrammetry, Remote Sensing, and Spatial Information Sciences*, 32(3/4W2), pp. 19–26.

Hussain, S., and B. Triggs, 2012. Visual recognition using local quantized patterns, *Proceedings of the European Conference on Computer Vision*, pp. 716–729.

Johnson, A. E., and M. Hebert, 1999. Using spin images for efficient object recognition in cluttered 3D scenes, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(5), pp. 433–449.

Kohli, P., L. Ladicky, and P.H.S. Torr, 2008. Robust higher order potentials for enforcing label consistency, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Krizhevsky, A., I. Sutskever, and G.E. Hinton, 2012. Imagenet classification with deep convolutional neural networks, *Advances in Neural Information Processing Systems*.

Ladicky, L., C. Russell, P. Kohli, and P. Torr, 2013. Associative hierarchical random fields, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6), pp. 1056–1077.

Ladicky´, L.,B. Zeisl, and M. Pollefeys, 2014. Discriminatively trained dense surface normal estimation, *Proceedings of the European Conference on Computer Vision*, pp. 468–484.

Lafarge, F., and C. Mallet, 2011. Building large urban environments from unstructured point data, *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1068–1075.

Lafarge, F., and C. Mallet, 2012. Creating large-scale city models from 3D-point clouds: A robust approach with hybrid representation, *International Journal of Computer Vision*, 99(1), pp. 69–85.

Lai, K., L. Bo, and D. Fox, 2014. Unsupervised feature learning for 3D scene labeling, *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 3050–3057.

Lalonde, J.-F., N. Vandapel, D. Huber, and M. Hebert, 2006. Natural terrain classification using three-dimensional ladar data for ground robot mobility, *Journal of Field Robotics*, 23(10), pp. 839–861.

Lawin, F.J., M. Danelljan, P. Tosteberg, G. Bhat, F.S. Khan, and M. Felsberg, 2017. Deep projective 3D semantic segmentation, *Proceedings of the International Conference on Computer Analysis of Images and Patterns*, LNCS 10424, Part I, Springer, Heidelberg, pp. 95–107.

LeCun, Y., B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, and L.D.Jackel, 1989. Backpropagation applied to handwritten zip code recognition, *Neural Computation*, 1(4), pp. 541–551.

Li, Y., S. Pirk, H. Su, C.R. Qi, and L.J. Guibas, 2016. FPNN: Field probing neural networks for 3D data, *Advances in Neural Information Processing Systems* D. D. Lee, M. Sugiyama, U.V. Luxburg, I. Guyon, and R. Garnett, editors), Vol. 29, pp. 307–315.

Lodha, S., E. Kreps, D. Helmbold, and D. Fitzpatrick, 2006. Aerial LiDAR Data Classification using Support Vector Machines (SVM), *Proceedings of the IEEE Third International Symposium on 3D Data Processing, Visualization, and Transmission*.

Long, J., E. Shelhamer, and T. Darrell, 2015. Fully convolutional networks for semantic segmentation, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints, *International Journal of Computer Vision*, 60(2), pp. 91–110.

Maas, H.-G., 1999. The potential of height texture measures for the segmentation of airborne laserscanner data, *Proceedings of the Fourth International Airborne Remote Sensing Conference and Exhibition/21st Canadian Symposium on Remote Sensing*, Vol. 1, pp. 154–161.

Malik, J., S. Belongie, T. Leung, and J. Shi, 2001. Contour and texture analysis for image segmentation, *International Journal of Computer Vision*, 43(1), pp. 7–27.

Manduchi, R., A.C., A. Talukder, and L. Matthies, 2005. Obstacle detection and terrain classification for autonomous off-road navigation, *Autonomous Robots*, 18, pp. 81–102.

Marton, Z.C., R.B. Rusu, and M. Beetz, 2009. On fast surface reconstruction methods for large and noisy point clouds, *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 3218–3223.

Maturana, D., and S. Scherer, 2015. Voxnet: A 3D convolutional neural network for real-time object recognition, *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 922–928.

Monnier, F., B. Vallet, and B. Soheilian, 2012. Trees detection from laser point clouds acquired in dense urban areas by a mobile mapping system, *ISPRS Archives of Photogrammetry, Remote Sensing, and Spatial Information Sciences*, I-3, pp. 245–250.

Montemerlo, M., and S. Thrun, 2006. Large-scale robotic 3-D mapping of urban structures. Experimental Robotics IX, *Springer Tracts in Advanced Robotics* (M. Ang and O. Khatib, editors),Vol. 21, Springer, Heidelberg, pp. 141–150.

Montoya, J., J.D.Wegner, L. Ladicky´ and K. Schindler, 2014. Mind the gap: Modeling local and global context in (road) networks, *Proceedings of the German Conference on Pattern Recognition, LNCS 8753*, Springer, Heidelberg, pp. 212–223.

Munoz, D., J.A. Bagnell, N. Vandapel, and M. Hebert, 2009a. Contextual classification with functional max-margin markov networks, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 975–982.

Munoz, D., N. Vandapel, and M. Hebert, 2009b. Onboard contextual classification of 3-D Point clouds with learned high-order Markov random fields, *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 2009–2016.

Niemeyer, J., F. Rottensteiner, and U. Soergel, 2014. Contextual classification of lidar data and building object detection in urban areas, *ISPRS Journal of Photogrammetry and Remote Sensing*, 87, pp. 152–165.

Niemeyer, J., J.D. Wegner, C. Mallet, F. Rottensteiner, and U. Soergel, 2011. Conditional random fields for urban scene classification with full waveform lidar data, *Photogrammetric Image Analysis, LNCS 6952*, Springer, Heidelberg, pp. 233–244.

Prokhorov, D., 2010. A convolutional learning system for object classification in 3-D lidar data, *IEEE Transactions on Neural Networks*, 21(5), pp. 858–863.

Qi, C.R., H. Su, K. Mo, and L.J. Guibas, 2017a. Pointnet: deep learning on point sets for 3D classification and segmentation, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 77–85.

Qi, C.R., L. Yi, H. Su, and L.J. Guibas, 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space, *Advances in Neural Information Processing Systems*.

Riegler, G., A.O. Ulusoy, and A. Geiger, 2017. Octnet: Learning deep 3D representations at high resolutions, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6620–6629.

Riemenschneider, H., A. Bo´ dis-Szomoru´ , J. Weissenberg, and L.Van Gool, 2014. Learning where to classify in multi-view semantic segmentation, *Proceedings of the European Conference on Computer Vision*, Springer, pp. 516–532.

Ronneberger, O., P. Fischer and T. Brox, 2015. U-net: Convolutional networks for biomedical image segmentation, *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, pp. 234–241.

Rottensteiner, F. and C. Briese, 2002. A new method for building extraction in urban areas from high-resolution lidar data, *ISPRS Archives of Photogrammetry, Remote Sensing, and Spatial Information Sciences* 34(3/A), pp. 295–301.

Rottensteiner, F., G. Sohn, M. Gerke, and J.D. Wegner, 2013. *ISPRS Test Project on Urban Classification and 3D Building Reconstruction, Technical Report*, ISPRS Working Group III/4 3D Scene Analysis.

Rottensteiner, F., G. Sohn, M. Gerke, J. Wegner, U. Breitkopf, and J. Jung, 2014. Results of the ISPRS benchmark on urban object detection and 3D building reconstruction, *ISPRS Journal of Photogrammetry and Remote Sensing*, 93, pp. 256–271.

Russakovsky, O., J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and L. Fei-Fei, 2015. Imagenet large scale visual recognition challenge, *International Journal of Computer Vision*, 115(3), pp. 211–252.

Rusu, R.B., Z.C. Marton, N. Blodow, A. Holzbach, and M. Beetz, 2009. Model-based and learned semantic object labeling in 3D point cloud maps of kitchen environments, *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3601–3608.

Serna, A., B. Marcotegui, F. Goulette, and J.-E.Deschaud, 2014. Paris-rue-madame database: A 3D mobile laser scanner dataset for benchmarking urban detection, segmentation and classification methods, *Proceedings of the 4th International Conference on Pattern Recognition, Applications and Methods*.

Shechtman, E., and M. Irani, 2007. Matching local self-similarities across images and videos, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Shotton, J., J. Winn, C. Rother, and A. Criminisi, 2006. Textonboost: Joint appearance, shape and context modeling for multiclass object recognition and segmentation, *Proceedings of the European Conference on Computer Vision, LNCS 3951*, Part I, Springer, Heidelberg, pp. 1–15.

Silberman, N., D. Hoiem, P. Kohli, and R. Fergus, 2012. Indoor segmentation and support inference from rgbd images, *Proceedings of the European Conference on Computer Vision*, Springer, pp. 746–760.

Simonyan, K., and A. Zisserman, 2014. Two-stream convolutional networks for action recognition in videos, *Advances in Neural Information Processing Systems*, pp. 568–576.

Simonyan, K., and A. Zisserman, 2015. Very deep convolutional networks for large-scale image recognition, *Proceedings of the International Conference on Learning Representations*.

Song, S., and J. Xiao, 2016. Deep sliding shapes for amodal 3D object detection in RGB-D images, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 808–816.

Song, S., S.P. Lichtenberg, and J. Xiao, 2015. Sun RGB-D: A RGB-D scene understanding benchmark suite, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 567–576.

Steder, B., R.B. Rusu, K. Konolige, and W. Burgard, 2010. Narf: 3D range image features for object recognition, *Proceedings of the Workshop on Defining and Solving Realistic Perception Problems in Personal Robotics at the IEEE/RSJ International Conference on Intelligent Robots and Systems*, Vol. 44.

Steder, B., R.B. Rusu, K. Konolige, and W. Burgard, 2011. Point feature extraction on 3D range scans taking into account object boundaries, *Proceedings of the IEEE International Conference on Robotics & Automation*, pp. 2601–2608.

Tatarchenko, M., A. Dosovitskiy, and T. Brox, 2017. Octree generating networks: Efficient convolutional architectures for high resolution 3D outputs, *arXiv* preprint arXiv:1703.09438.

Tombari, F., S. Salti, and L. Di Stefano, 2010. Unique signatures of histograms for local surface description, *Proceedings of the European Conference on Computer Vision*, Springer, pp. 356–369.

Torralba, A., R. Fergus, and W.T. Freeman, 2008. 80 million tiny images: A large data set for nonparametric object and scene recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11), pp. 1958–1970.

Torralba, A., K. Murphy, and W. Freeman, 2004. Sharing features: Efficient boosting procedures for multiclass object detection, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Vallet, B., M. Bre´dif, A. Serna, B., Marcotegui, and N. Paparoditis, 2015. Terramobilita/iqmulus urban point cloud analysis benchmark, *Computers & Graphics* (49), pp. 126–133.

Vandapel, N., D. Huber, A. Kapuria, and M. Hebert, 2004. Natural terrain classification using 3-D ladar data, *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 5117–5122.

Weinmann, M., B. Jutzi, and C. Mallet, 2013. Feature relevance assessment for the semantic interpretation of 3D point cloud data, *ISPRS Archives of Photogrammetry, Remote Sensing, and Spatial Information Sciences* II-5(W2), pp. 313–318.

Weinmann, M., S. Urban, S. Hinz, B. Jutzi, and C. Mallet, C., 2015. Distinctive 2D and 3D features for automated large-scale scene analysis in urban areas, *Computers & Graphics*, 49, pp. 47–57.

Wu, Z., S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, 2015. 3D shapenets: A deep representation for volumetric shapes, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1912–1920.

Yan, W., A. Shaker, and N. El-Asjmawy, 2015. Urban land cover classification using airborne LiDAR data: A review, *Remote Sensing of Environment*, 158, pp. 295–310.

Yao, W., S. Hinz, and U. Stilla, 2011. Extraction and motion estimation of vehicles in single- pass airborne lidar data towards urban traffic analysis, *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(3), pp. 260–271.

Zeiler, M.D., 2012. Adadelta: an adaptive learning rate method, *arXiv preprint arXiv*:1212.5701.

Zhuang, Y., G. He, H. Hu, and Z. Wu, 2015a. A novel outdoor scene-understanding framework for unmanned ground vehicles with 3D laser scanners, *Transactions of the Institute of Measurement and Control*, 37(4), pp. 435–445.

Zhuang, Y., Y. Liu, G. He, and W. Wang, 2015b. Contextual classification of 3D laser points with conditional random fields in urban environments, *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3908–3913.

Zwicker, M., H. Pfister, J. Van Baar, and M. Gross, 2001. Surface splatting, *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, ACM, pp. 371–378.

Spatiotemporal Change Detection Based on Persistent Scatterer Interferometry: A Case Study of Monitoring Building Changes

C. H. Yang, B. K. Kenduiywo, and U. Soergel

Abstract

Persistent scatterer interferometry (PSI) detects and analyses PS points from multitemporal SAR images for scene deformation monitoring. We propose a novel technique to identify disappearing and emerging PS points, which are regarded as building changes in cities. A spatiotemporal analysis is implemented as both spatial position and occurrence time are obtained for each change point. We first estimate each pixel's temporal coherences in different image subsets. Computed from temporal coherences, a change index sequence is introduced to quantify each pixel's probabilities of being change points at different times. All pixels' change indices are then utilized to extract change points by a global, automatic, and statistical-based scheme. We then eliminate blunders by a spatial filtering. Finally, the occurrence times of the change points are detected based on the temporal variation in their change index sequences. We implement a simulated data test to validate and assess our method. Using TerraSAR-X images, our real data test successfully recognizes steady, disappearing, and emerging buildings in Berlin, Germany within 2013.

Introduction

The continuous rise in population and economic growth has led to urbanization across the world coming along with frequent building changes, e.g., construction, in a built-up environment. Monitoring such changes is important for city management, urban planning, updating of cadastral maps, etc. (Gamba, 2013; Marin *et al.*, 2015). Remote sensing offers a fast and cost-effective mapping of large areas compared with conventional field surveys. Particularly, spaceborne synthetic aperture radar (SAR) sensors provide radar images captured rapidly over vast areas at fine spatiotemporal resolution. The SAR sensors are weather independent and have a day-and-night vision ability, which guarantees images with a high temporal density. These capabilities make SAR suitable for monitoring events.

Many time series analysis methods using multitemporal SAR images have been proposed for urban monitoring. For instance, persistent scatterer interferometry (PSI) (Costantini *et al.*, 2008; Crosetto *et al.*, 2005 and 2016; Ferretti *et al.*, 2000, 2001, and 2011; Hooper *et al.*, 2004; Kampes, 2006) detects persistent scatterer (PS) points, which are characterized by strong, stable, and coherent radar signals throughout a SAR image sequence. In principle, PSI eventually derives for each PS a set of attributes, such as temporal coherence, line-of-sight (LoS) velocity (mm/year level), topography height, and geographic position. These attributes are then used for scene monitoring. A signal sequence of a PS point is modeled to maintain coherence during the entire acquisition period of a SAR image stack. Accordingly, a scene of interest covered with PS points is assumed to be steady and free of big changes. For example, PSI works well in monitoring of built-up cities because the regular and stationary substructures of buildings cause high PS density. However, if the substructures or even entire buildings disappear due to construction, the corresponding semi-PS points are discarded in the initial screening of PSI processing. In other words, big changes cannot be revealed by common PSI.

Some previous works (Ansari *et al.*, 2014; Brcic and Adam, 2013; Ferretti *et al.*, 2003; Novali *et al.*, 2004) aim at detecting semi-PS points, which disappear or emerge at arbitrary times, by looking for abrupt amplitude changes of pixels along SAR image stack. Indeed, the amplitude-based thresholding methods (Adam *et al.*, 2003; Crosetto *et al.*, 2003; Ferretti *et al.*, 2001; Lyons and Sandwell, 2003; Werner *et al.*, 2003) are commonly used to choose PS candidates, e.g., by means of amplitude dispersion (Ferretti *et al.*, 2001). Here, high and stable amplitudes indicate potential PS points. However, we might miss those low-amplitude PS points even their signals are permanently coherent and stable. Consequently, low-amplitude semi-PS points cannot be recognized either. To avoid such loss, an alternative strategy (Hooper *et al.*, 2004) exploits temporal coherence, an indicator of temporal phase stability, for PS identification. This strategy is adapted and extended in our method for change detection.

In this study, we propose spatiotemporal change detection based on PSI to detect disappearing and emerging semi-PS points along with their occurrence times. The term "spatiotemporal" refers to changes that occur over geographical space and are detected over time. We distinguish and label these two scenarios as disappearing big change (DBC) and emerging big change (EBC) points. To begin with, multitemporal SAR images are divided into several subsets by a sequence of break dates (*bd*). The temporal coherence of each pixel in an image set is estimated using a standard PSI processing. Temporal coherence is modeled to be proportional to phase stability and thus serves as an indicator of a PS point. The key idea of our approach is to derive a change index sequence for each pixel from its temporal coherence estimates spanning different periods. An automatic thresholding is applied to the change indices to determine the final change points. We then eliminate blunders by a spatial filtering, which is a crucial element for a spatiotemporal analysis. Finally, we check the evolution of a change index sequence to identify the probable

C. H. Yang and U. Soergel are with the Institute for Photogrammetry, University of Stuttgart, Geschwister-Scholl-Str. 24D, 70174 Stuttgart, Germany (yang@ifp.uni-stuttgart.de).

B. K. Kenduiywo is with the Department of Geomatic Engineering and Geospatial Information Systems, Jomo Kenyatta University of Agriculture and Technology, P.O. Box 62000-00200, Nairobi, Kenya.

Photogrammetric Engineering & Remote Sensing
Vol. 84, No. 5, May 2018, pp. 309–328.
0099-1112/18/309–328

doi: 10.14358/PERS.84.5.309

occurrence date of a change point if any. Before getting into the methodological details, we first introduce the PSI processing used in this study.

Persistent Scatterer Interferometry

A time series of N complex SAR images acquired using the same system parameters is a prerequisite for PSI. Among them, N–1 interferograms are generated based on a master image, which is optimally chosen under small baseline constraint (Berardino *et al.*, 2002; Lanari *et al.*, 2004) to diminish temporal and geometric de-correlations. The interferometric phases are modeled as:

$$\varphi^{int}(x) = \varphi^{int}_{res\_topo}(x) + \varphi^{int}_{mot}(x) + \varphi^{int}_{APS}(x) + \varphi^{int}_{noise}(x) \quad (1)$$

where x denotes the pixels in the interferograms indexed by int $[1, N-1] \in \mathrm{N}$. The four phase components are interpreted as follows. Topographic error $\varphi^{int}_{res\_topo}$, which is caused by residual Δh after flattening (Crosetto *et al.*, 2016; Hanssen, 2001; Kampes, 2006) is formulated as:

$$\varphi^{int}_{res\_topo}(x) = \frac{4\pi \cdot B^{int}_{\perp}}{\lambda \cdot R(x) \cdot \sin[\theta(x)]} \cdot \Delta h(x) \quad (2)$$

where $B_{\perp}$, λ, R, and θ indicate perpendicular baseline, wavelength of SAR signal, slant range between SAR antenna and ground target of x, and look angel of SAR signal. The motion-induced phase φ^{int}_{mot} is expressed as:

$$\varphi^{int}_{mot}(x) = \frac{4\pi}{\lambda} \cdot B^{int}_{T} \cdot v(x) \quad (3)$$

where $B_{\perp}$ and v denote temporal baseline and LoS velocity, respectively. Both residual topographic height Δh and LoS velocity v are regarded as two unknowns to be solved for each pixel. In contrast, φ^{int}_{APS}, which is caused by atmospheric phase screen (APS) (Zebker *et al.*, 1997), and φ^{int}_{noise}, a sum of de-correlation noise stemming from thermal noise, processing errors, inaccuracy of orbital parameters, temporal and geometric de-correlations, etc. (Kampes, 2006), are treated as phase disturbance.

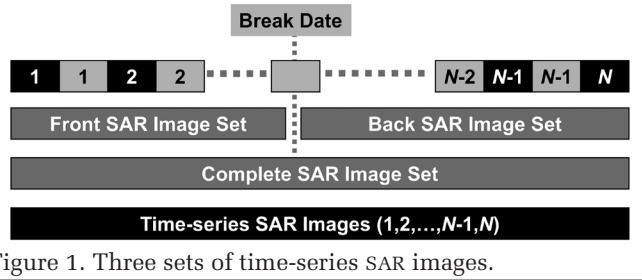


Figure 1. Three sets of time-series SAR images.

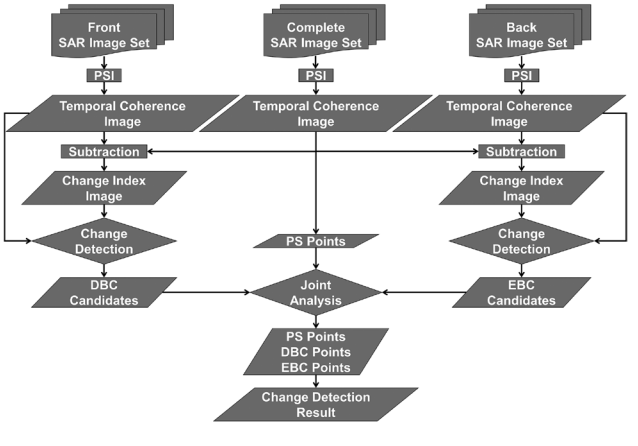


Figure 2. Flowchart of single-break-date change detection scheme. Persistence, disappearance, and emergence scenarios are dedicated to extracting PS, DBC, and EBC points using complete, front, and back SAR image sets, respectively.

The optimal estimates of the pixels' residual topographic error $\Delta\hat{h}$ and LoS velocity $\hat{v}$ are determined using periodogram searching (Ferretti *et al.*, 2001)

$$\underset{\Delta h(x)\ \&\ v(x)}{\operatorname{argmax}} \left\{ \gamma_T(x) = \left| \frac{1}{N-1} \cdot \sum_{int=1}^{N-1} \exp j\left[\varphi^{int}_{o}(x) - \varphi^{int}(x)\right] \right| \right\} \quad (4)$$

where argmax denotes an arguments of the maxima, γ_T indicates temporal coherence, and φ^{int}_{o} are phase observations. However, the APS disturbance in φ^{int}_{o} contaminates the estimates and must be excluded in the second searching to improve the precision. After the first searching, the residual phases calculated as:

$$\varphi^{int}_{res}(x) = \varphi^{int}_{o}(x) - \hat{\varphi}^{int}_{res\_topo}(x) - \hat{\varphi}^{int}_{mot}(x) \quad (5)$$

are mainly composed of APS-induced phases φ^{int}_{APS} and noise-induced phases φ^{int}_{noise}. Suppose that the APS phases are spatially correlated but temporally uncorrelated, the spatiotemporal filtering (Ferretti *et al.*, 2000 and 2001) is applied to the residual phases $\varphi^{int}_{\mathrm{res}}$ to derive the APS phases $\hat{\varphi}^{int}_{APS}$. The periodogram searching is iterated after subtracting $\hat{\varphi}^{int}_{APS}$ from the phase observations φ^{int}_{o}. As a result, the estimates' precisions and temporal coherences should be improved; if not, the whole procedure must be exhaustively checked to solve the problems.

Generally, the coherence estimates in interferometry tend to be biased towards the optimistic side in case of low number of samples (i.e., N) and low true coherences (Bamler and Hartl, 1998; Touzi *et al.*, 1999). Conventionally, for medium-resolution SAR sensors like ERS or Envisat, a minimum stack of 15 images is considered a prerequisite for PSI (Crosetto *et al.*, 2016). For high-resolution sensors even less images can be sufficient (Bovenga *et al.*, 2012). However, to be on the safe side, we later use a least 16 TerraSAR-X images in PSI processing and a rather high threshold of temporal coherence for PS selection. We believe our parameters are sufficient for PSI implementation given an urban scene.

Finally, pixels are selected as PS points if their temporal coherences fulfill a specified threshold while the unselected pixels are discarded. However, those discarded pixels might include change points, which cannot be detected at this stage by PSI. To overcome this shortcoming, we can now turn to our novel methodology.

Methodology

Single-Break-Date Scheme

We first illustrate the change detection scheme subject to a single break date that big changes occur before or after. More accurate change times can be distinguished later when multi-break-date scheme is conducted. Complete, front, and back SAR image sets are defined from an image series for use in this scheme (Figure 1). The complete set consists of all images. The front and back sets comprise the images taken before and after the break date, respectively. Our aim is to find change points that exist as PS points in the front set but then disappear in the back set and vice versa.

The flowchart (Figure 2) is composed of the persistence, disappearance, and emergence scenarios, in which the

complete, front, and back sets are mainly involved to detect PS, DBC, and EBC points. These three image sets are processed by a standard PSI procedure to generate three temporal coherence images (real values in the range from 0 to 1). The temporal coherence image of the complete set is subtracted from those of the front and back sets to obtain two change index images ranging from −1 to 1. High change indices signify probable change events. The change indices are used to extract the change points. They are jointly analyzed with the PS points, which are selected in the persistence scenario, to exclude two types of outliers. First, if a change label coincide with a PS label, the pixel is labeled as PS considered to be more reliable. Second, two different change labels cannot happen simultaneously. In case of such conflict, the corresponding points are regarded as void points same as the remaining unlabelled pixels. Finally, the PS and change points are combined for further analysis.

Change Index

We introduce a change index defined by temporal coherence to quantify each pixel's probability of being a change point. Our approach assumes that the temporal coherence estimates of a PS point in complete, front, and back sets are approximately the same. In contrast, the temporal coherence of a change point in a front or back set is higher than that in a complete set, which suffers from coherence loss due to big change. Based on these assumptions, the change indices of a pixel x in disappearance (CI^D) and emergence (CI^E) scenarios are calculated by:

$$\underset{[-1\ ,+1]\in \mathrm{R}}{CI^{D}(x)} = \gamma_T^F(x) - \gamma_T^C(x) \tag{6}$$

$$\underset{[-1\ ,+1]\in \mathrm{R}}{CI^{E}(x)} = \gamma_T^B(x) - \gamma_T^C(x) \tag{7}$$

where $\gamma_{\mathrm{T}}^{\mathrm{C}}$, $\gamma_{\mathrm{T}}^{\mathrm{F}}$, and $\gamma_{\mathrm{T}}^{\mathrm{B}}$ denote temporal coherences in complete, front, and back sets. A pixel is more likely to be a DBC or EBC point when CI^D or CI^E more tends towards 1, respectively. In contrast, change indices approximating 0 indicate potential PS points.

Automatic Thresholding

Based on change indices, we design a global, automatic, and statistical-based thresholding method to extract change points. Given that no big change occurs, a change index distribution over PS points in disappearance or emergence scenario is assumed to follow a Gaussian distribution:

$$N\left(CI_{PS}(x)\,|\,\mu,\sigma\right) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{-CI_{PS}^2(x)}{2\sigma^2}\right) \tag{8}$$

The temporal coherences of each PS point in complete, front, and back sets should be approximately identical. Thus, the mean μ of change indices calculated by Equations 6 or 7 is anticipated to be 0. The standard deviation indicates an overall PS quality considering two PSI scenarios. For instance, high-quality PS points give rise to a narrow and tall curve of change index distribution characterized by $\mu \approx 0$ and small σ. In contrast to PS points, a change index distribution over change points does not conform to Gaussian because the big changes substantially and arbitrarily alter their temporal coherences between complete, front, and back sets. The significant difference between change index distributions over PS and change points inspired us to design a thresholding technique.

An original change index distribution (Figure 3) is modeled as the sum of a Gaussian curve plus a non-symmetric probability distribution function (large right tail). The Gaussian curve originates from change indices of major PS points; the right tail is caused by high change indices of minor change points. The goal is to automatically determine an optimal change index threshold to extract the change points with as few false points as possible. Our idea is to fit a Gaussian curve characterized by only the PS points without the large right tail. This fitted curve assists in delimiting a dividing line, i.e., a change index threshold, to separate the large right tail from the original curve.

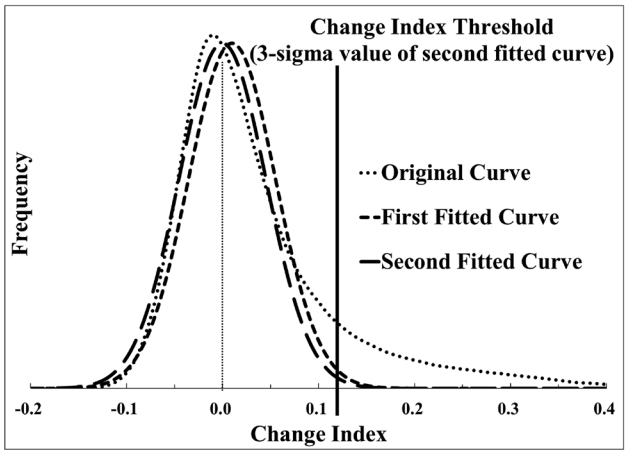


Figure 3. Examples of original, first fitted, and second fitted curves of change index distribution. 3-sigma value indicates triple standard deviation in second fitted curve.

The Gaussian curve is modeled as:

$$Frequency\left(CI(x)\right) = H_G\exp\left(-\left(\frac{CI(x)-M_G}{S_G}\right)^2/2\right) \tag{9}$$

where H_G, M_G, and S_G denote height, mean, and standard deviation. There are two iterations in the curve fitting. The first iteration fits all of the elements (change indices and frequencies) of the original curve (Figure 3) to the Gaussian curve model (Equation 9). The first fitted curve is then estimated by least squares (Teunissen, 2000). However, this curve skews to the right because of the influence from the elements of the large right tail. To diminish this skew, we perform a second iteration: only the elements of the partial first fitted curve, which are bounded within triple the standard deviation (3-sigma value), are used to estimate the second fitted curve. Consequently, the skew vanishes as the second fitted curve is more centered on 0 compared with the first one. We conclude that the second fitted curve is mainly characterized by the PS points and is less affected by the change points.

During the curve fitting process, the 3-sigma value accounts for 99.7 percent of the elements of the first fitted curve. This value is set based on a trade-off between the influences of the PS and change points on the second fitted curve. For example, increasing the sigma value yields a more precise Gaussian bell shape of the second fitted curve characterized by more PS points; however, this curve should skew more to the right because the influence of change points is also augmented. This skewness might result in a non-optimal change index threshold, which leads to more false points. We suggest

using a 3-sigma value as a balance according to our experience and convincing results.

Finally, the change index threshold (Figure 3) is calculated as triple the standard deviation of the second fitted curve. In principle, the threshold delimits the Gaussian curve to the left side and the large right tail to the other side as far as possible. The change points are then extracted if their change indices are larger than the threshold. The intention of using 3-sigma value as a strict threshold is to avert as many false change points as possible.

Limitations

There are two limitations for the single-break-date scheme. First, detection of big changes is dependent on a pre-set break date, which can be manually set for specific interests. This requirement restricts the applicability particularly when *a priori* knowledge of scene changes is unavailable. Second, accurate occurrence times of big changes are lacking as they are only known to disappear or emerge after a break date. We combine several single-break-date results to overcome these two limitations.

Multi-Break-Date Scheme

The multi-break-date scheme (Figure 4) demands a set of single-break-date results, which are subject to a series of break dates, as input. For each pixel, two sequences, i.e., change indices and initial point labels (PS, DBC, EBC, or void), have been determined thus far. A joint analysis is applied to each pixel's initial point labels to decide its final label. A pixel is labeled as PS if all of its initial point labels are PS. In contrast, a pixel initially labeled as DBC or EBC is finally labeled as such. However, in case both change labels coincide, the points are rather labeled as void to avoid contradiction. Once PS and change points are confirmed, the remaining unlabelled pixels are considered to be void points too. An outlier filtering is then utilized to remove potential false points. Afterwards, the change date of each change point is detected from the break date series based on the temporal variation in its change index sequence. In the end, the PS and change points along with the change dates are combined to illustrate the spatiotemporal results.

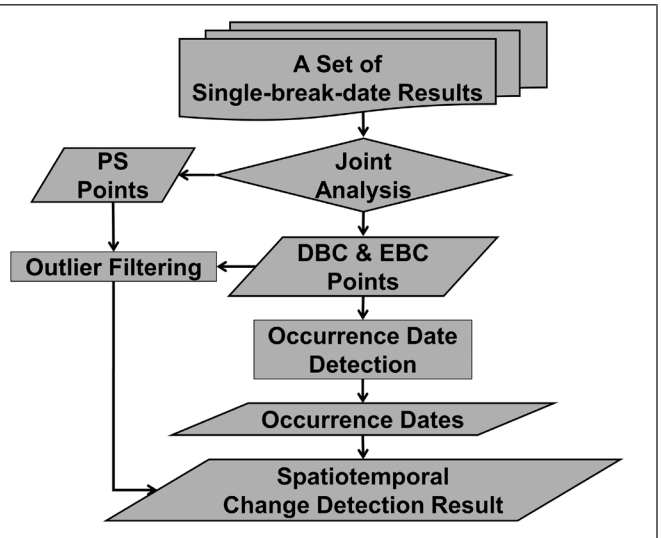


Figure 4. Flowchart of multi-break-date change detection scheme.

Spatial Outlier Filtering

Two outlier types are described below along with their removal strategies using sliding window operation.

- Different change labels within a window are removed considering that identical change labels are assumed to be clustered as scenes of a certain size.
- A single point except void label within a window is removed as its reliability cannot be inspected by comparing with neighbors.

Change Date Detection

We design an automatic way to detect change points' occurrence dates (an interval between two successive image acquisitions). Consider a DBC point, we simulate a typical evolution of change indices calculated by (Equation 6) (Figure 5). The temporal coherence over the entire stack is degraded because the object disappears on some date. The temporal coherence stays high as long as the front image set is still in the period before the disappearance, leading to large and constant change indices. In contrast, after the object disappears, starts to drop because we add more and more images, which no longer contain the related signals, to the front image set. Consequently, the change indices is gradually decaying as well. The same argumentation holds for EBC points. In summary, detecting a disappearance or emergence date is equivalent to locating the turning point in a change index sequence. To do so, we introduce a simple geometric algorithm. First, a horizontal line 1 extends from the sequence beginning to the left. Starting from the terminal of line 1, a straight line 2 is drawn to the end of the sequence. The turning point featuring the longest distance, line 3, to line 2 can then be detected. Finally, the corresponding break date is regarded as the disappearance date. We adopt a similar process to detect emergence dates of EBC points as well.

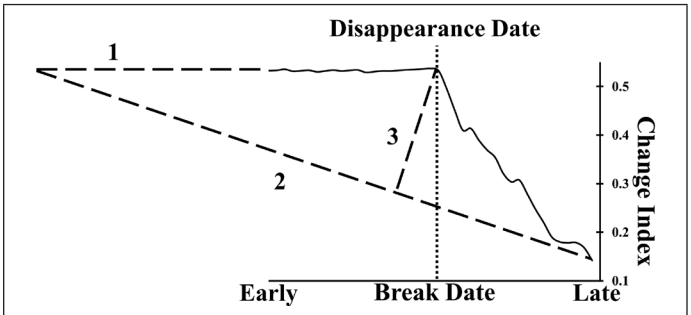


Figure 5. Change index sequence pertaining to disappearance.

Simulated Data Test

Simulation Procedure

We simulated a time series of M interferograms where PS, DBC, EBC, and void points are randomly generated. First of all, we assign a stochastic constant phase φ_{cons} to the simulated phases $\varphi_{sim}^{int}\left[-\pi\ ,\pi\right) \in R$ of each pixel x:

$$\varphi_{sim}^{int}(x) = \varphi_{cons}(x), int = \left[1\ ,M\right] \tag{10}$$

Gaussian phase noise φ_{n}^{int} ($\mu = 0$, $\sigma = [-\pi, \pi) \in \mathrm{R}$) is then added to the simulated phases:

$$\varphi_{sim}^{int}(x) = \varphi_{cons}(x) + \varphi_{n}^{int}(x) \tag{11}$$

The temporal coherences of $\varphi_{\text{sim}}^{int}(x)$ are calculated by;

$$\underset{[0\ ,1]\in \mathrm{R}}{\gamma_T(x)} = \left|\frac{1}{M}\cdot\sum\nolimits_{\text{int}=1}^{M}\exp\left(j\bullet\varphi_{sim}^{int}(x)\right)\right| \tag{12}$$

Pixels are selected as PS points if their temporal coherences fulfill a specified threshold. Among them, in case a DBC or EBC point disappears or emerges right after a *bd*, a series of irregular phases $\varphi_{irr}^{int^D}$ or $\varphi_{irr}^{int^E}$ is added to its simulated phases as:

$$\varphi_{sim}^{int}(x) = \varphi_{cons}(x) + \varphi_{n}^{int}(x) + \varphi_{irr}^{int^D}(x), int^D = [bd+1, M] \quad (13)$$

$$\varphi_{sim}^{int}(x) = \varphi_{cons}(x) + \varphi_{n}^{int}(x) + \varphi_{irr}^{int^E}(x), int^E = [1, bd] \quad (14)$$

respectively. Finally, those pixels without any label are regarded as void points.

Simulated Data

We simulated a scene with numerous big changes occurring randomly in time to analyze the performance of our technique. For this purpose, we generated 80 time-series interferograms (500 × 500) containing 58 percent PS, 17 percent DBC, 17 percent EBC, and 8 percent void points. A temporal coherence threshold of 0.8 is used for PS selection. The disappearance and emergence dates are evenly distributed from *bd*: 31 to 51. One example of a PS point (Figure 6a) shows that its change index sequence is close to 0 with a standard deviation of 0.004. Consequently, all of the initial point labels are PS and thus lead to final PS label. The examples of DBC and EBC points with an occurrence date of *bd*: 41 are illustrated in Figures 6b and 6c. Some initial void labels occur for two reasons. First, their temporal coherences do not fulfill the threshold of 0.8 necessary for change candidates in disappearance or emergence scenario. Second, their change indices are too low to be labeled as change points. Nevertheless, such void labels among initial point labels are ignored to determine a final point label. The turning points of the change index sequences correspond to the occurrence of *bd*: 41. This correspondence conforms to our assumption to detect events' dates.

Accuracy Assessment

The confusion matrix (Table 1) demonstrates that the overall accuracy is 99 percent and all of the producer's and user's accuracies are better than 99 percent. The only errors are the change points falsely labeled as PS, which result in producer's accuracies of 99 percent for both change labels and a user's accuracy of 99 percent for PS. The temporal coherences of these change points in the complete set are still above the threshold by chance. Such erroneous instances might happen especially when the number of SAR images in a front or back set takes only a small part of the entire image stack. The accuracy of the estimated dates of events is assessed by the comparison with the reference (Figure 7). In each disappearance or emergence example, we calculate the mean of the estimated occurrence dates for each reference date. The correlation coefficient of 0.999 demonstrates that our method can detect events' dates with a considerably high accuracy.

Table 1. Confusion matrix.

		Reference				
		PS	**DBC**	**EBC**	**Void**	**Sum**
Result	**PS**	145544	110	93	0	145747
	DBC	0	41647	0	0	41647
	EBC	0	0	41713	0	41713
	Void	0	0	0	20893	20893
	Sum	145544	41757	41806	20893	250000

	PS	**DBC**	**EBC**	**Void**
Producer's Accuracy	100%	99%	99%	100%
User's Accuracy	99%	100%	100%	100%
Overall Accuracy	99%			

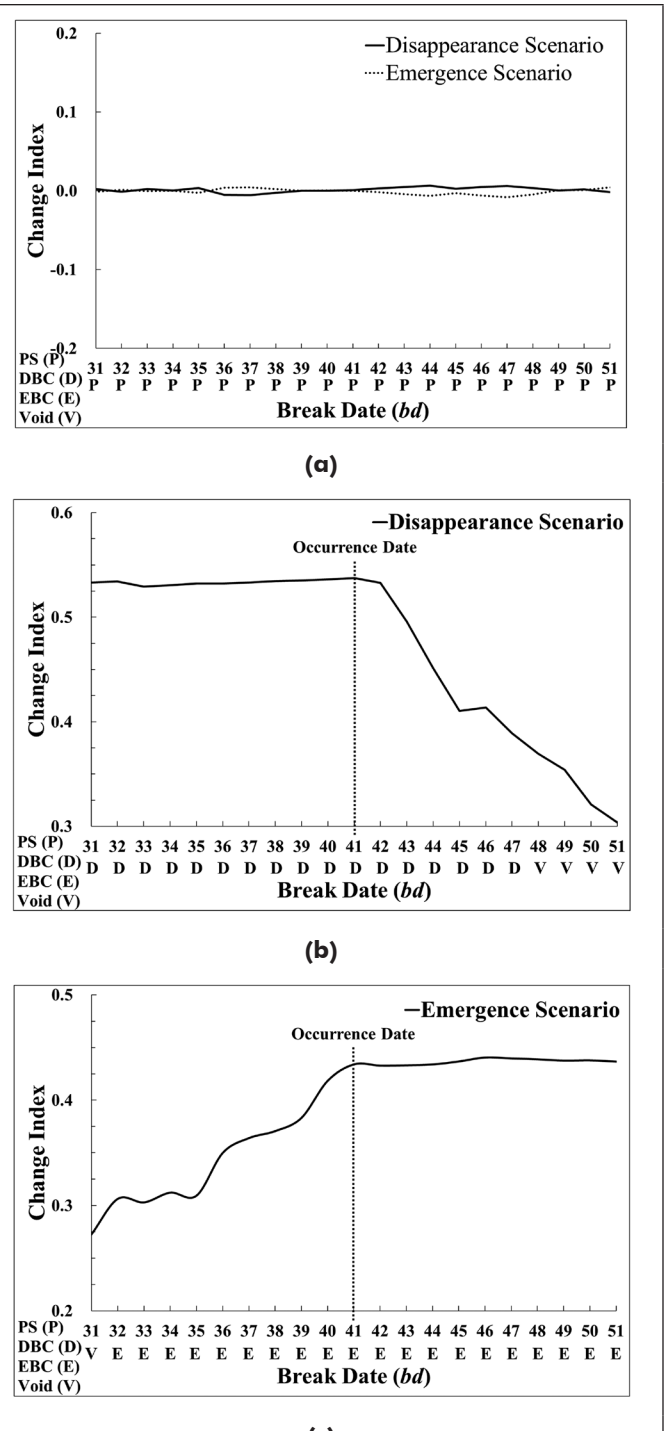


Figure 6. Simulated change index sequences of (a) PS, (b) DBC, and (c) EBC points.

Real Data Test

Data

The study area (Figure 8) covering the city center in Berlin, Germany, shows many bright clusters of strong signals on buildings that appear to be potential PS and change points. We adopted forty TerraSAR-X images (Table 2) acquired in High Resolution Spotlight mode from 27 October 2010 to 04 September 2014. All of the images were precisely co-registered and resampled into 5,000 × 5,000 grid (ground resolution: 1 m). The thirteen break dates (*bd*: 16 to 28) were chosen

Table 2. TerraSAR-X images and break date setup.

Acquisition Dates of TSX Images			
2010/10/27	2011/08/31	2013/08/26 *bd=21*	2014/02/07
2010/11/18	2011/10/03	2013/09/17 *bd=22*	2014/03/01
2011/01/23	2011/12/30	2013/09/28 *bd=23*	2014/03/23
2011/02/14	2012/01/10	2013/10/20 *bd=24*	2014/05/06
2011/03/08	2012/02/01	2013/10/31 *bd=25*	2014/05/28
2011/03/30	2012/02/12 *bd=16*	2013/11/22 *bd=26*	2014/06/19
2011/06/04	2013/06/21 *bd=17*	2013/12/03 *bd=27*	2014/07/11
2011/06/15	2013/07/13 *bd=18*	2013/12/25 *bd=28*	2014/08/02
2011/07/18	2013/07/24 *bd=19*	2014/01/05	2014/08/24
2011/08/20	2013/08/15 *bd=20*	2014/01/16	2014/09/04

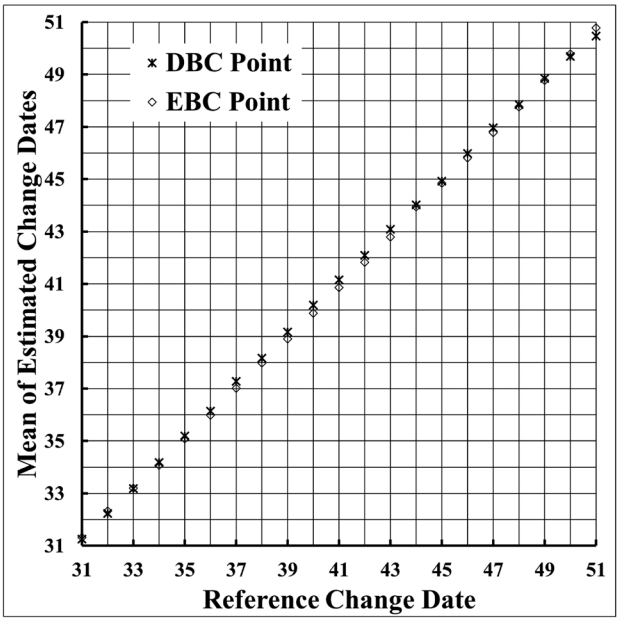
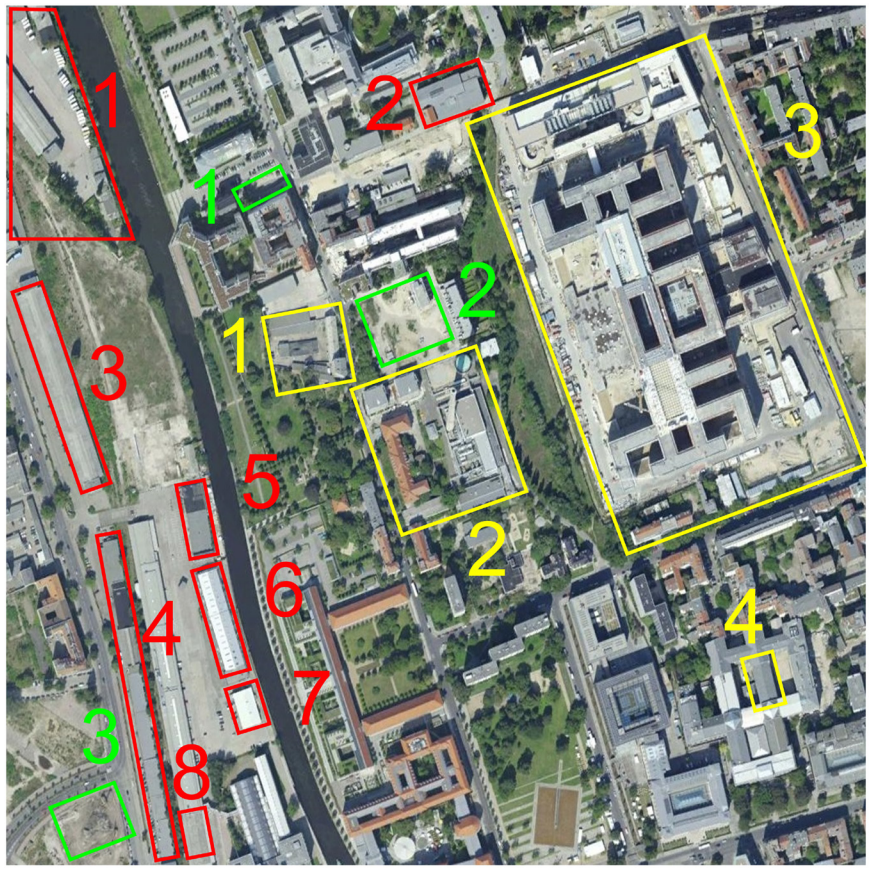
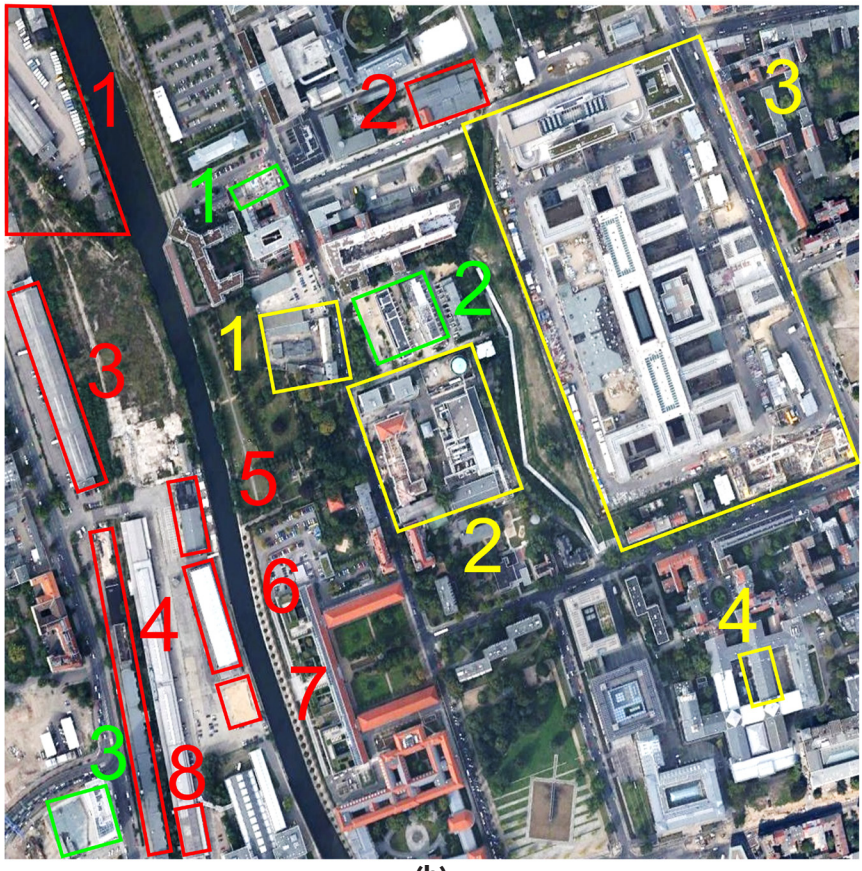


Figure 7. Mean of estimated events' dates versus reference. Correlation coefficient of 0.999 for both change types.

Figure 8. Mean TerraSAR-X image over study area. Patch 1 is used for in-depth analysis.

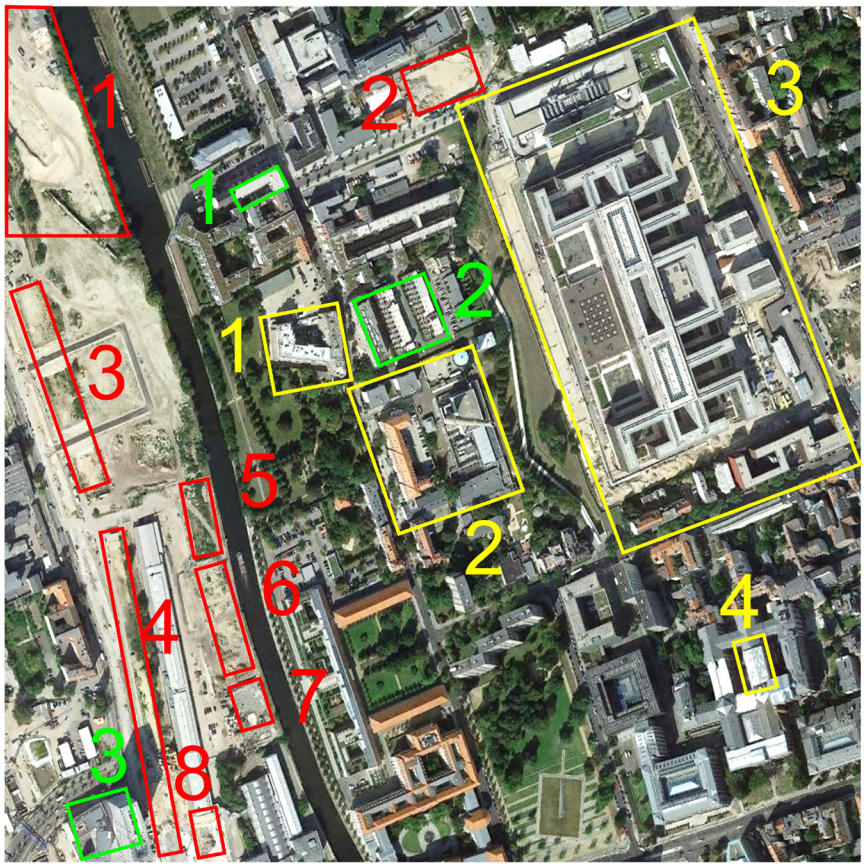


(a)



(b)

Figure 9. Aerial images (©Google Earth) over patch 1 (Figure 8) around Berlin Central Station acquired on (a) 12 September 2010, (b) 20 May 2012, and (c, *see next page*) 05 September 2014. Building change: red, disappearance area; green, emergence area; yellow, hybrid area.



(c)

Figure 9 *continued*. Aerial images (©Google Earth) over patch 1 (Figure 8) around Berlin Central Station acquired on (a) 12 September 2010, (b) 20 May 2012, and (c) 05 September 2014. Building change: red, disappearance area; green, emergence area; yellow, hybrid area.

to explore the ground changes within 2013. We compared our results with three Google™ Earth's aerial images (ground reference) taken on 12 September 2010, 20 May 2012, and 05 September 2014.

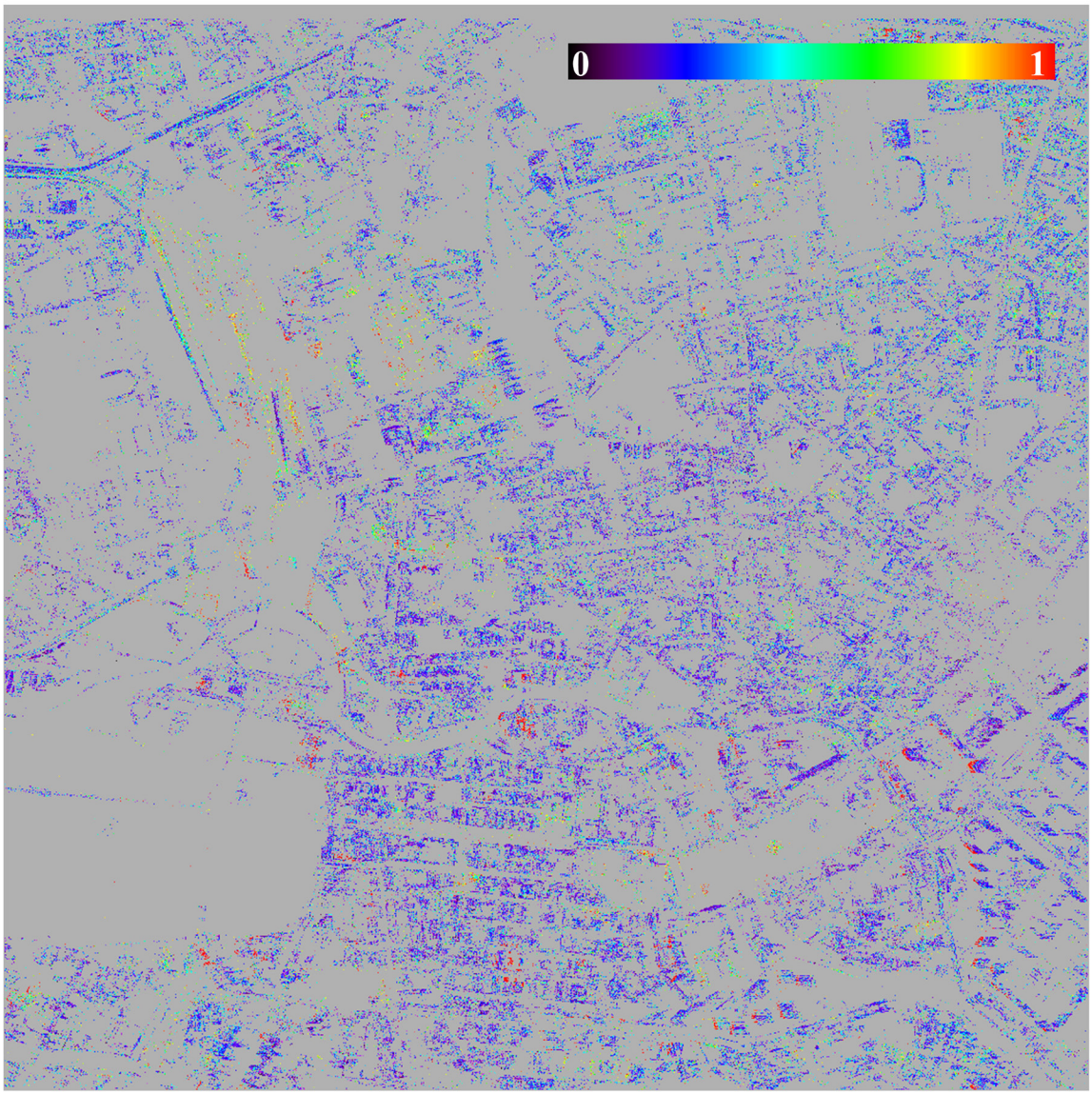
Our analysis will focus on patch 1 (Figure 9) around Berlin Central Station where various construction activities have taken place. The buildings in disappearance areas (red) 1 to 8 were demolished after September 2010. Therefore, we expect dense DBC points to be detected inside these areas. We identify some new buildings in emergence areas 1 (green) to 3 where EBC points are anticipated on the newly-built substructures. The building changes in hybrid areas 1 to 4 (yellow) cannot be attributed entirely to either disappearance or emergence. Two different buildings are present in hybrid area 1 on 12 September 2010 and 05 September 2014, respectively. A similar case is also found in hybrid area 4. Our method is able to tell what happened in 2013. We observe renovation activities in hybrid area 2 where different point labels are supposed to be detected. The headquarters of the Federal Intelligence Service was built in hybrid area 3 starting in October 2006. Our aim is to monitor the construction progress in detail to find out where substructures were removed or constructed and when these changes occurred.

Single-Break-Date Result

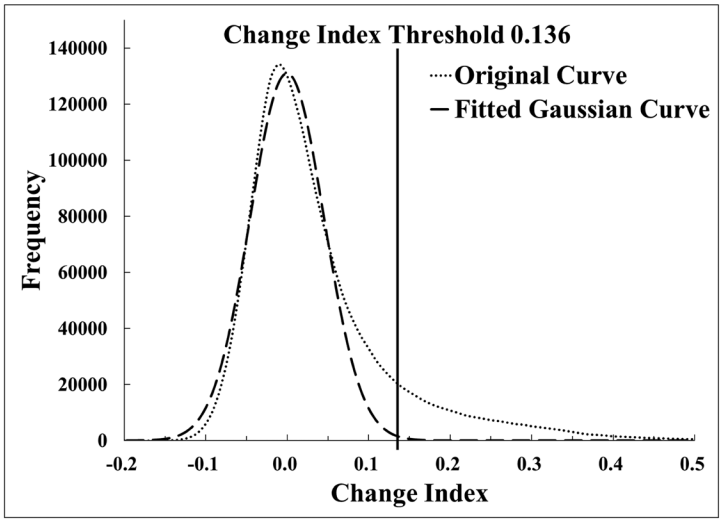
We introduce a single-break-date result subject to the break date between 12 February 2012 and 21 June 2013 (*bd*: 16, Table 2) to demonstrate how to detect change points based on change indices. The change indices of the PS points, which pertain to the front set in the disappearance scenario, indicate their probabilities of being DBC points (Figure 10a). Those buildings highlighted by high change indices were more likely to be demolished close to or after the break date. The original curve of the change index distribution consists of a Gaussian curve and a large right tail, which are supposed to be caused by PS and DBC points, respectively (Figure 10b). The change index threshold of 0.136, calculated as triple the standard deviation of the fitted Gaussian curve, is adopted for DBC detection. Similarly, EBC points can be extracted using the change indices computed in the emergence scenario (Figure 11). We can recognize the disappearing and emerging structures in the change detection result (Figure 12). Overall, the result within patch 1 (Figure 13) conforms to the ground reference (Figure 9). However, the accurate occurrence times are still unknown, which is the main limitation of the single-break-date scheme. To overcome this weakness, we then consider the multi-break-date result in the next section.

Multi-Break-Date Result

The spatiotemporal change detection result (Figure 14) reveals where the changed structures are along with their occurrence times. Compared with the single-break-date result (Figure 12), the numbers of the DBC and EBC points are increased by 111 percent and 276 percent. Examples of DBC and EBC points are illustrated in Figures 15 and 16. Their characteristics are consistent with our methodological assumptions. The results in patch 1 (Figure 17) correspond to the ground reference (Figure 9). The buildings in disappearance areas 1 and 3 to 7 were demolished by the middle break date, followed by those in areas 2 and 8. For example, in area 2, the building

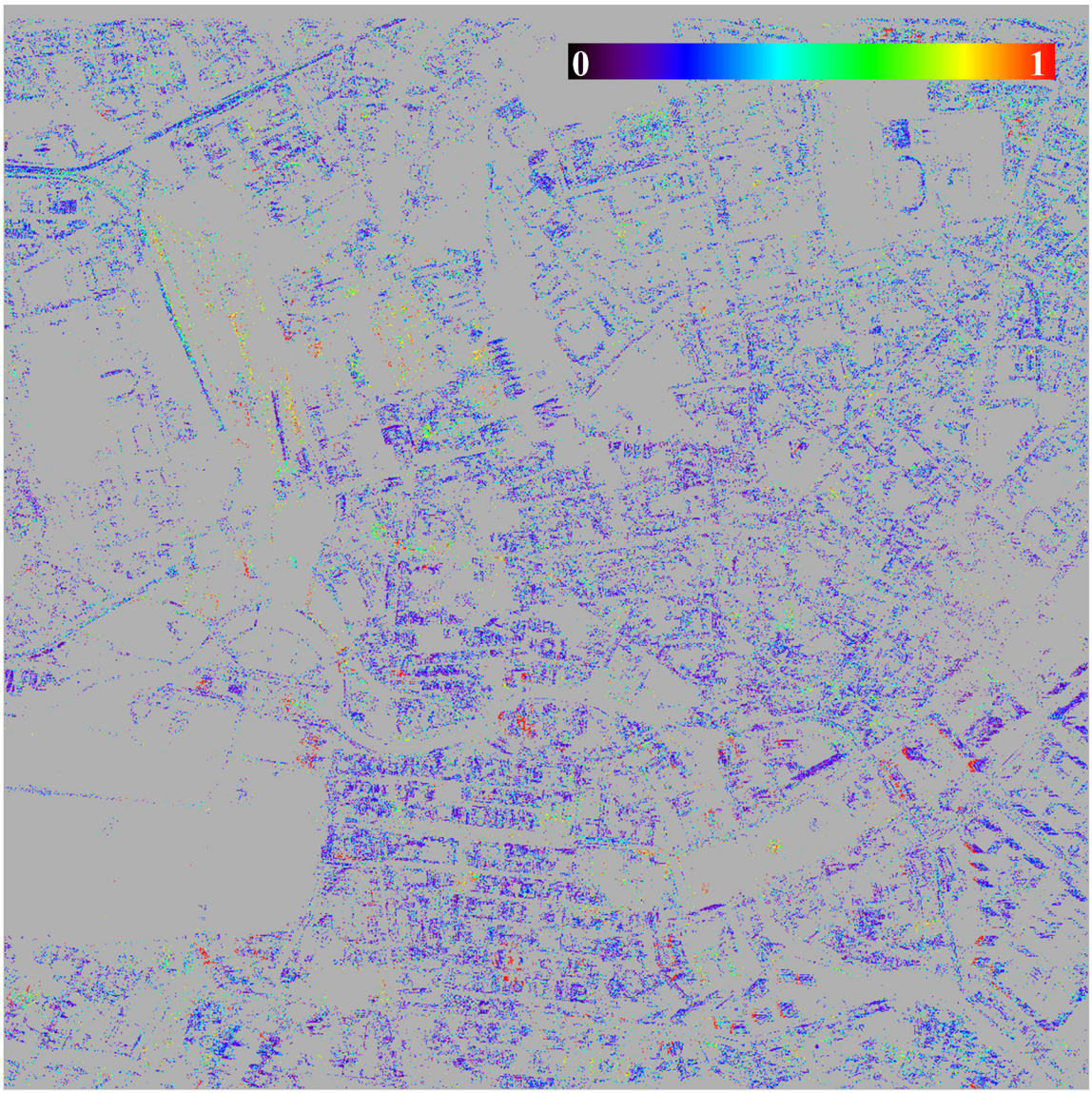


(a)

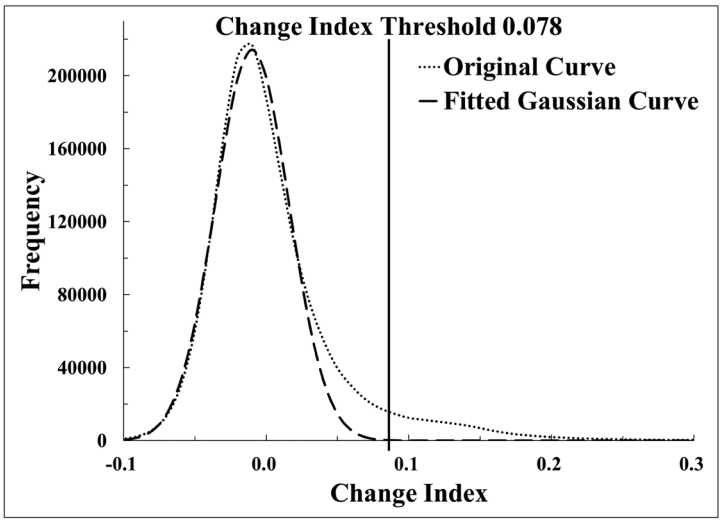


(b)

Figure 10. Disappearance scenario: (a) change index image; and (b) original and fitted curves of change index distribution.



(a)



(b)

Figure 11. Emergence scenario: (a) change index image; and (b) original and fitted curves of change index distribution.

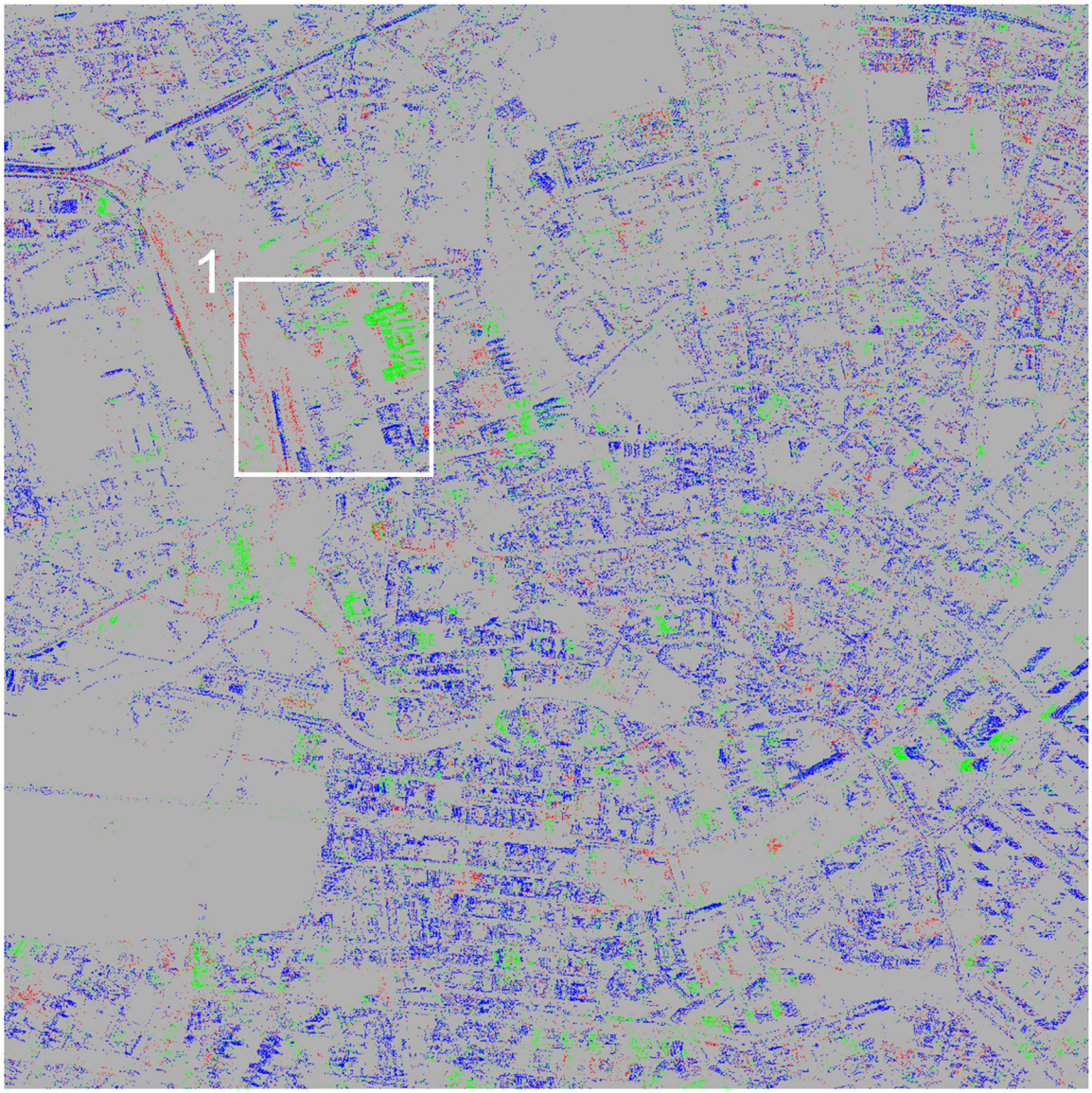


Figure 12. Change detection result: steady, disappearing, and emerging structures represented by PS (blue, 49868/km^2), DBC (red, 9185/km^2), and EBC (green, 14765/km^2) points. Patch 1 is used for in-depth analysis.

existed on 20 May 2012 and was completely demolished before 05 September 2014. This fact verifies our finding from the result that the substructures disappeared gradually after the middle break date. Emergence area 1 reveals a new building erected in early 2013. The ground reference shows that part of this building had been constructed on 20 May 2012 and the remainder should be completed soon after. In emergence area 2, two new apartments on the right side appeared on 20 May 2012, which leads to the EBC points found on the early dates. In contrast, the EBC points on the third new apartment, which appeared on 05 September 2014, occurred later. Emergence area 3 shows that only parts of the new office building had been constructed by the end of 2013. An early-built building is present in hybrid area 1. Hybrid area 2 provides the renovation information that certain substructures were removed or added during the second half of the break dates. The results in hybrid area 3 illustrate an example of construction progress monitoring with detailed change information. The building-shaped pattern of the dense EBC points indicates that the main building structure had been constructed in early 2013 and then other substructures were added to it over time. In addition, we also observe some disappearing substructures. They are considered to be the construction materials on the facades and foundations that were removed or covered in the early stage. Finally, the building in hybrid area 4 was demolished after the middle break date.

Comparison with Ratio Change Detection

We compared our technique with the conventional ratio change detection (Rignot and van Zyl, 1993) to underline its advantages. For this analysis, we chose the two images acquired on 12 February 2012 and 21 June 2013 (Table 2), respectively. During this period many construction activities took place. The data were converted into intensity images, which were then divided by each other, and the result was expressed in dB. The ratio values of unchanged objects concentrate at 0 as a Gaussian distribution; in contrast, those of changed objects tend towards positives and negatives. Finally, we adopted Otsu thresholding (Otsu, 1979) to extract changes.

We focus our comparison and analysis on patch 1 (Figure 17). The ratio image (Figure 18a) manifests the potential changes highlighted in black and red. Among the detected

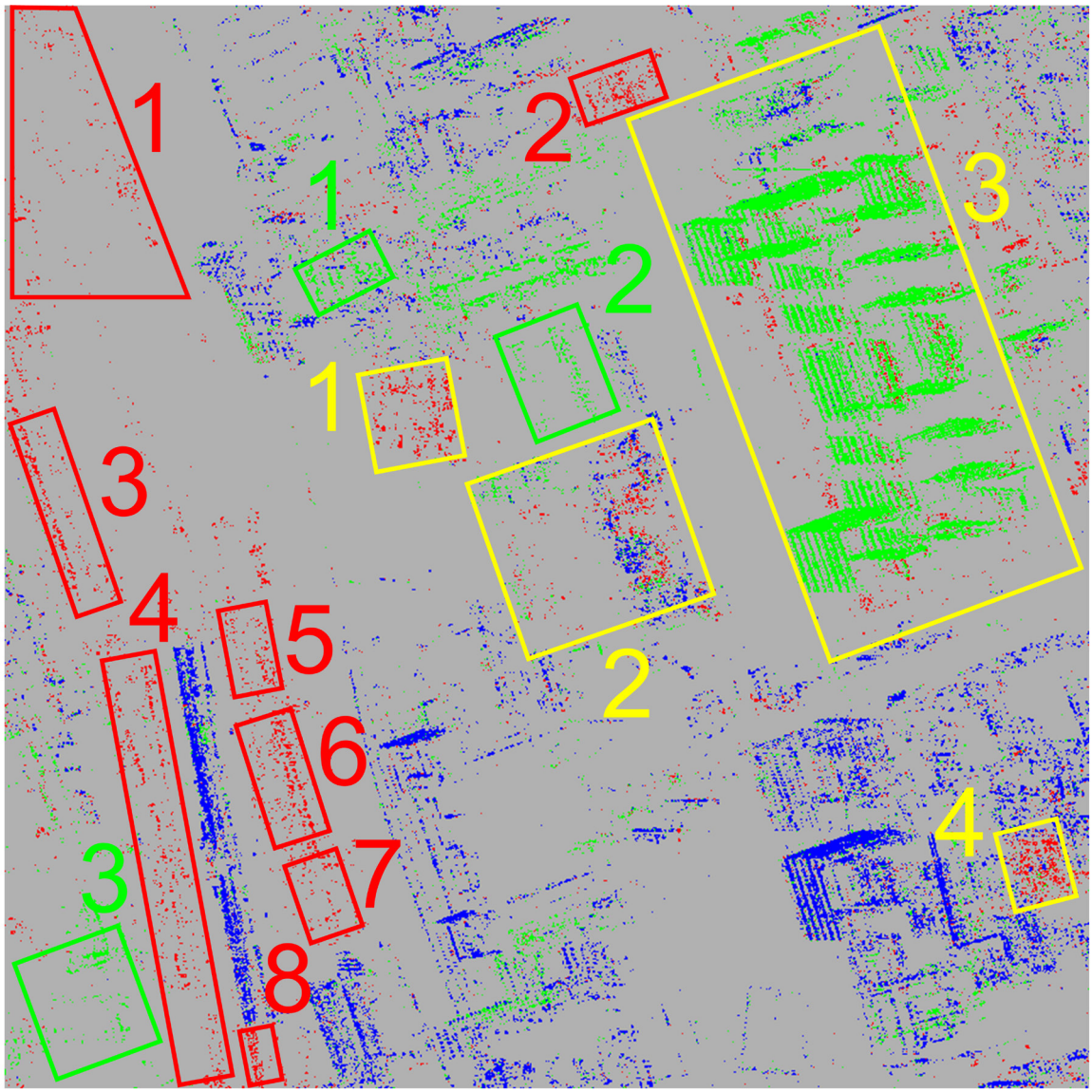


Figure 13. Change detection result within patch 1 (Figure 12).

changes (Figure 18b) we can identify indeed clusters which are caused by change of interest, i.e., construction activities. However, in addition we observe all over the scene salt-and-pepper responses above threshold, which seem to stem from issues like speckle or noise. Both types of results are kind of superimposed in the detected changes, leading to difficulty in interpretation. In a second experiment, we preprocessed the intensity images by speckle suppression based on refined Lee filter (Lee, 1981) of size 5 × 5. As a result, the changes due to construction (Figure 19) can be more clearly identified now; however, false alarms, in particular noise, still dominate the result. Finally, we applied multi-looking by a factor of 100 to diminish both speckle and noise in the intensity images. As a result, most false alarms were eliminated (Figure 20). Nevertheless, we certainly also lose spatial details. Last, but not least, different change types cannot be easily distinguished from a ratio image.

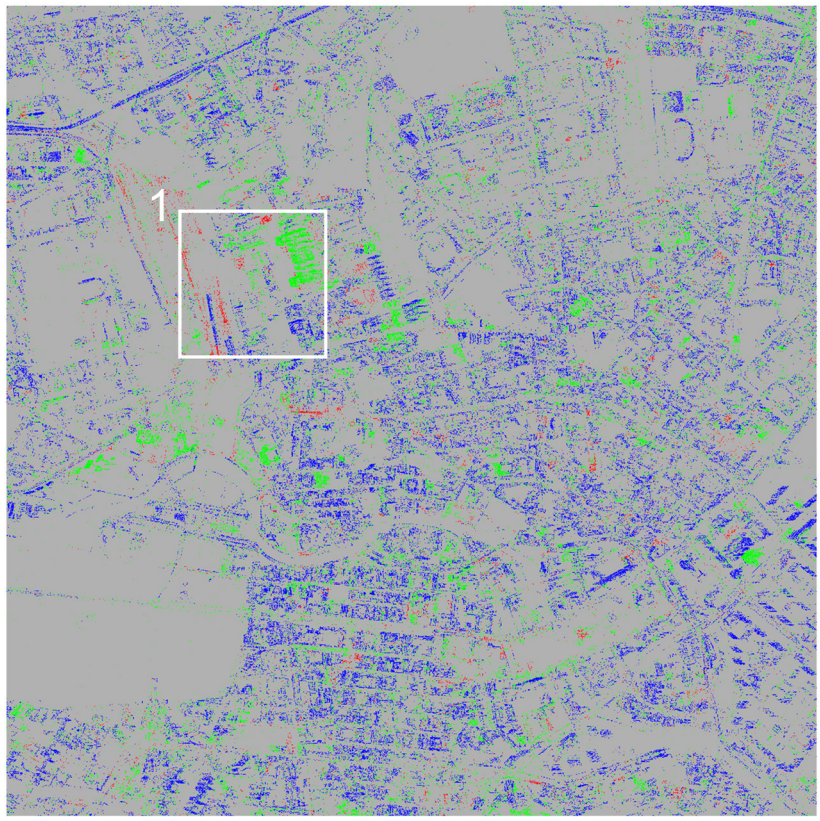
Those problems mentioned above exist not only for ratio change detection but also for other incoherent approaches. The strategy of our method looks for disappearance and emergence of PS points. Therefore, we only deal with building changes and disregard other change types, speckle, and noise. Our technique also benefits from high-resolution SAR images to detect detailed changes in a more accurate way.

Comparison with Amplitude-Based Semi-PS Method

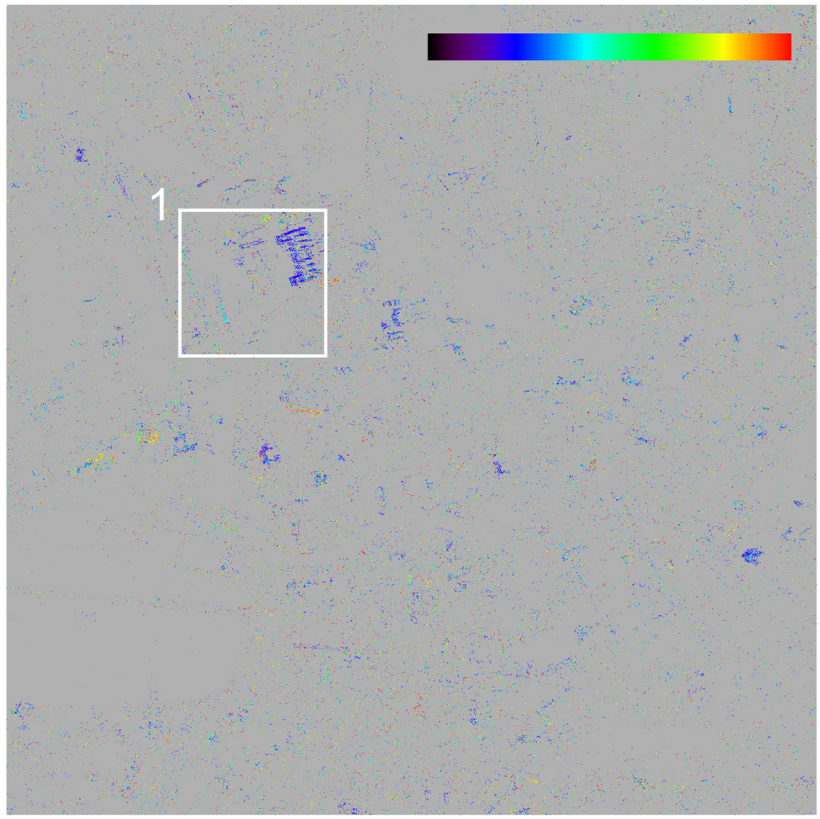
We implemented the amplitude-based method (Ferretti *et al.*, 2003) for comparison with our approach. The overall result (Figure 21) displays the change points detected among the thirteen break dates (Table 2). The number of these change points is remarkably less than those (Figure 14a) identified by our new technique, which leads to 223 percent and 498 percent increases for DBC and EBC points. Comparing the patch 1 results (Figures 17a and 22) shows that the amplitude-based result loses partial details or even all of the change structures. For example, most of the EBC points are missing in hybrid area 2, which fails to convey the renovation activity. In summary, our approach working on phase information has proven capable of detecting more change points, i.e., more comprehensive information.

Conclusions

We propose a spatiotemporal change detection technique capable of detecting spatial big changes along with their occurrence times by using multitemporal SAR images as single-source data. The simulated data test yields accurate spatiotemporal changes. The overall accuracy is 99 percent and all of the producer's



(a)



(b)

Figure 14. Spatiotemporal change detection result. Patch 1 is used for in-depth analysis. (a) Steady, disappearing, and emerging structures represented by PS (blue, 49868/km^2), DBC (red, 19343/km^2), and EBC (green, 55509/km^2) points. (b) Disappearance and emergence dates: black to red, earliest to latest in 2013.

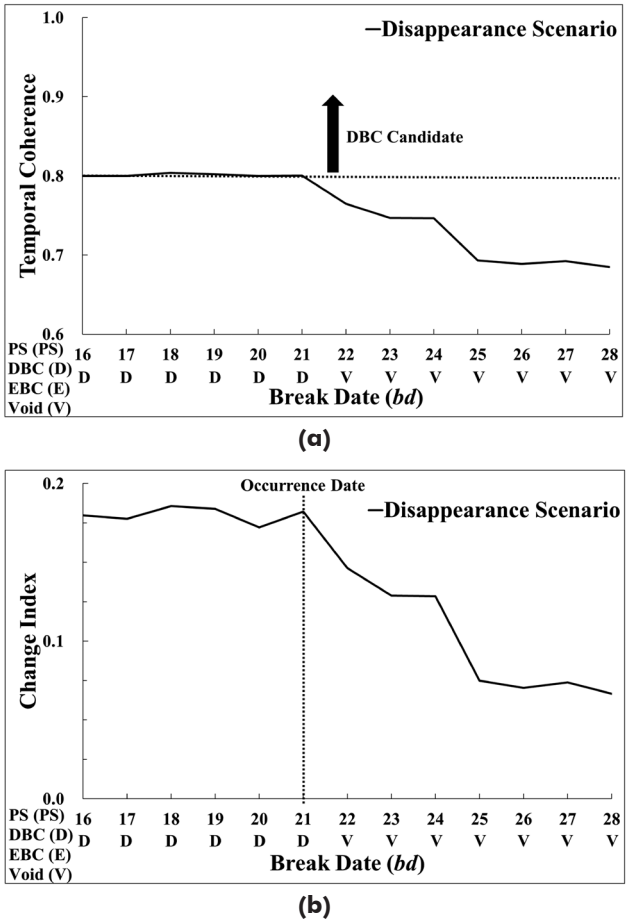


Figure 15. Example of DBC point's (a) temporal coherence sequence, and (b) chance index sequence along with initial point labels.

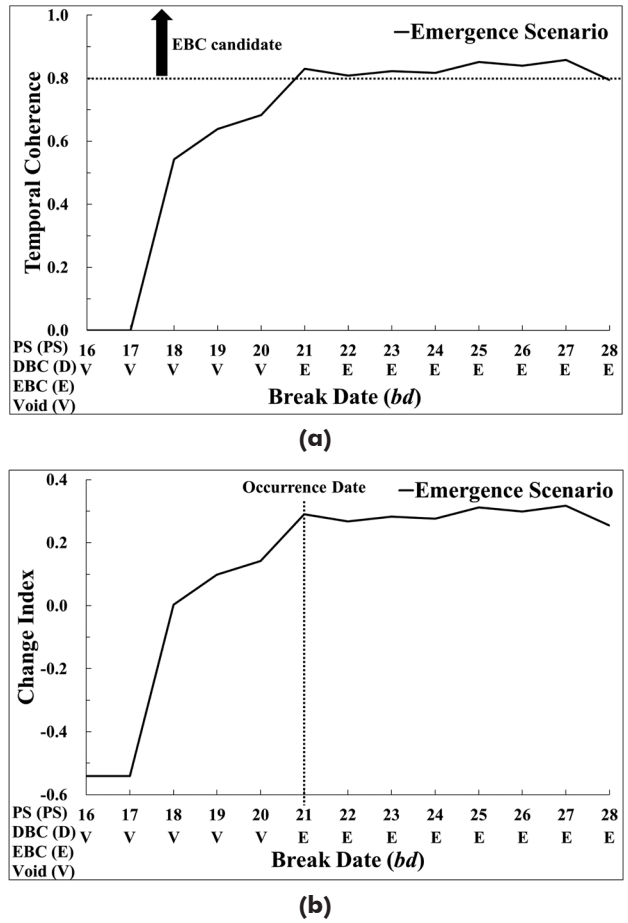


Figure 16. Example of EBC point's (a) temporal coherence sequence, and (b) chance index sequence along with initial point labels.

and user's accuracies are better than 99 percent. The means of the estimated change dates were compared with the reference dates. The correlation coefficient of 0.999 demonstrates that our method can detect change dates with a considerably high accuracy. We successfully detect the steady, disappearing, and emerging buildings in Berlin, Germany, within 2013 by using a stack of forty TerraSAR-X images. The spatiotemporal change detection result provides detailed information to interpret various change events. Compared with the ratio change detection, our technique is able to target only structural changes and disregard other change types, speckle, and noise. We also demonstrate that more change points can be detected by our method than those by the amplitude-based semi-PS approach.

The proposed approach is particularly suitable for urban monitoring. For example, we can separate destroyed buildings and damaged substructures due to natural disasters, such as earthquake, from other construction activities, which took place before. In the aftermath of the disaster, we could monitor subsequent reconstruction as well. In addition, we can easily adapt our approach based on an alternative time-series SAR interferometry for different purposes. For instance, our technique working on distributed scatterers enables changes on natural objects, such as rocky terrain, to be detected. In practice, the detail level of spatiotemporal changes depends on spatiotemporal resolution of SAR images. In this study, we are able to explore the meter-resolution substructures that disappeared or emerged within, at best, 11 days.

We have four plans for the future. To reinforce our technique, we will test and optimize the statistical model of change index distribution. A possible alternative is a mixture of two Gaussian distributions caused by PS and change points. Second, we will seek a more robust analysis, e.g., maximum likelihood estimator, to identify change points. Third, integration of complementary data, e.g., SAR amplitude images, will be considered to further improve the performance. Finally, we will investigate the accuracy of change detection with respect to temporal coherence threshold, number of images, statistical model, etc.

Acknowledgments

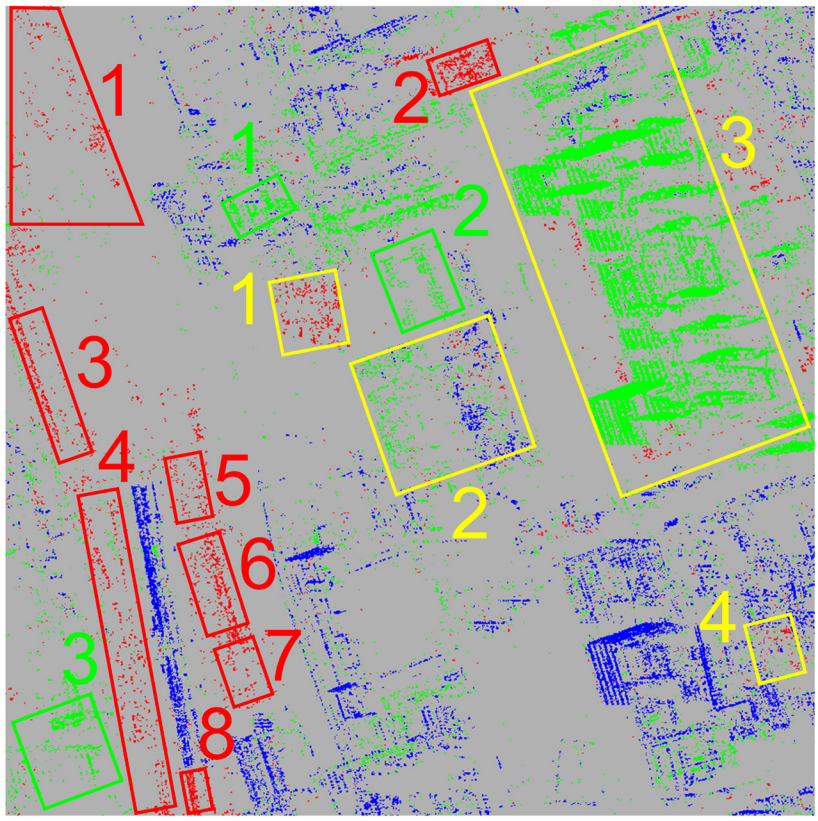
The authors would like to appreciate the anonymous reviewers whose sincere comments largely improves the paper quality. The TerraSAR-X images used in this study were provided by DLR.

References

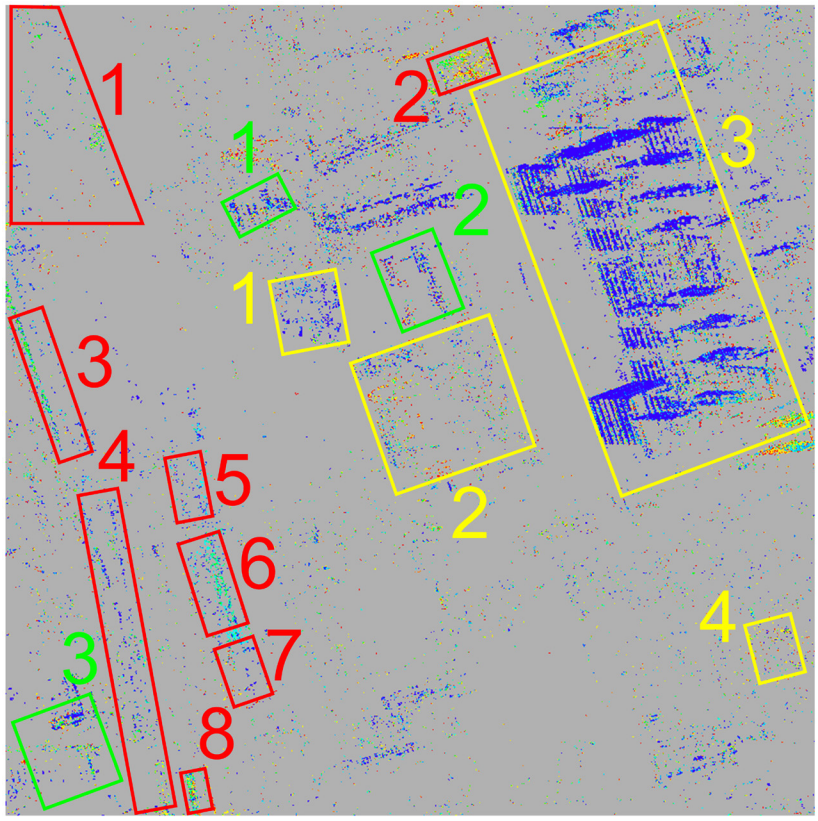
Adam, N., B. Kampes, M. Eineder, J. Worawattanamateekul, and M. Kircher, 2003. The development of a scientific permanent scatterer system, *Proceedings of the ISPRS Workshop*, Hannover, Germany.

Ansari, H., N. Adam, and R. Brcic, 2014. Amplitude time series analysis in detection of persistent and temporal coherent scatterers, *Proceedings of the 2014 IEEE International Geoscience and Remote Sensing Symposium*, Quebec, Canada, 2213–2216.

Bamler, R., and P. Hartl, 1998. Synthetic aperture radar interferometry, *Inverse Problems*, 14(4):R1–R54.

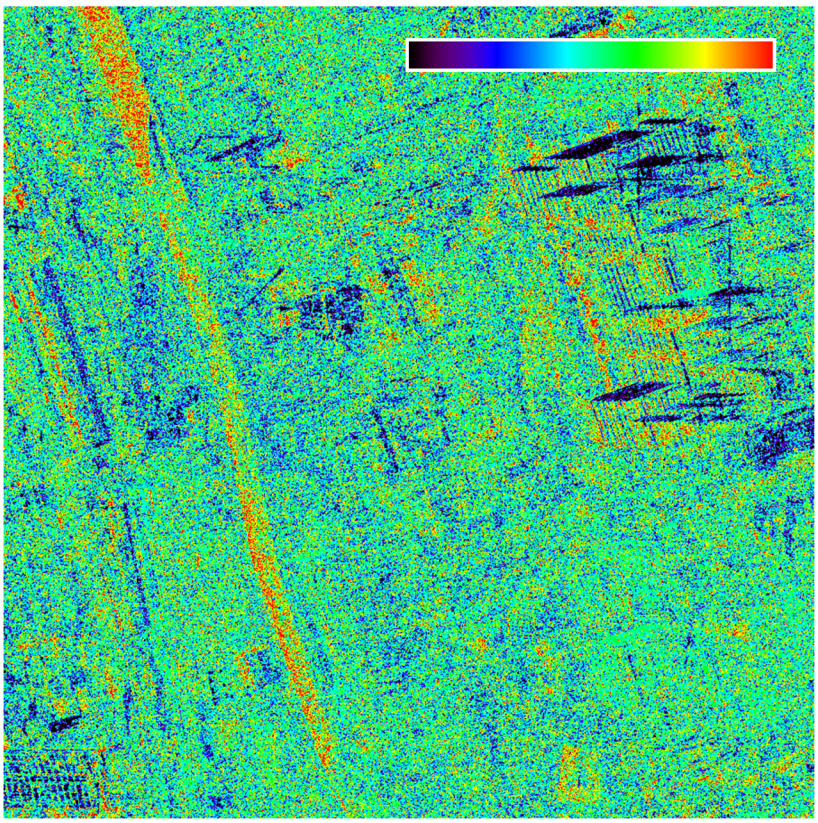


(a)

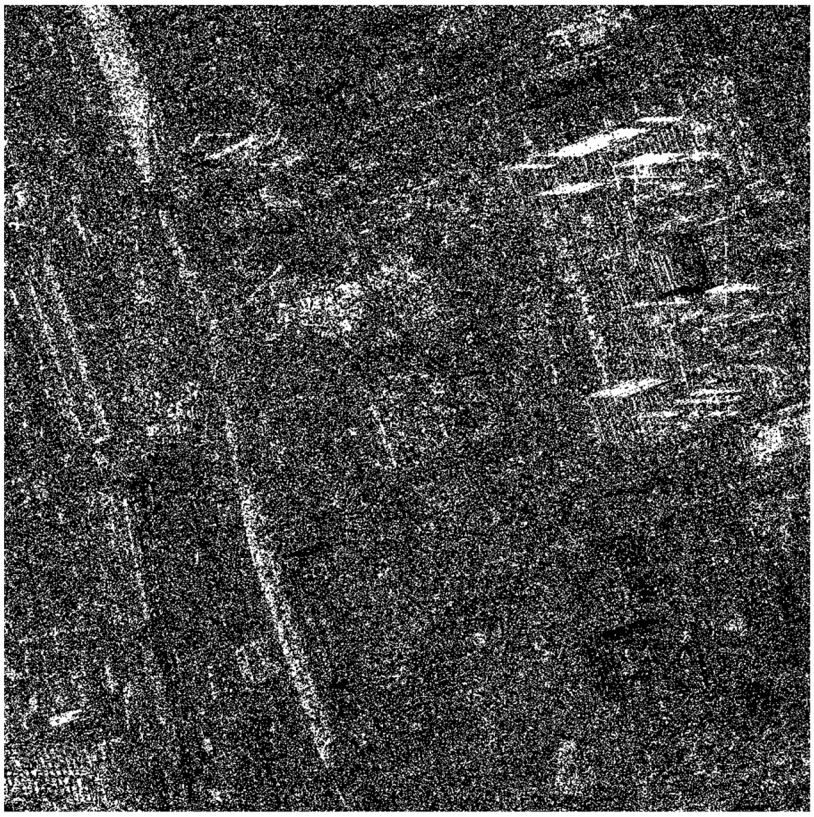


(b)

Figure 17. Spatiotemporal change detection result within patch 1 (Figure 14). Building change: red, disappearance area; green, emergence area; yellow, hybrid area.

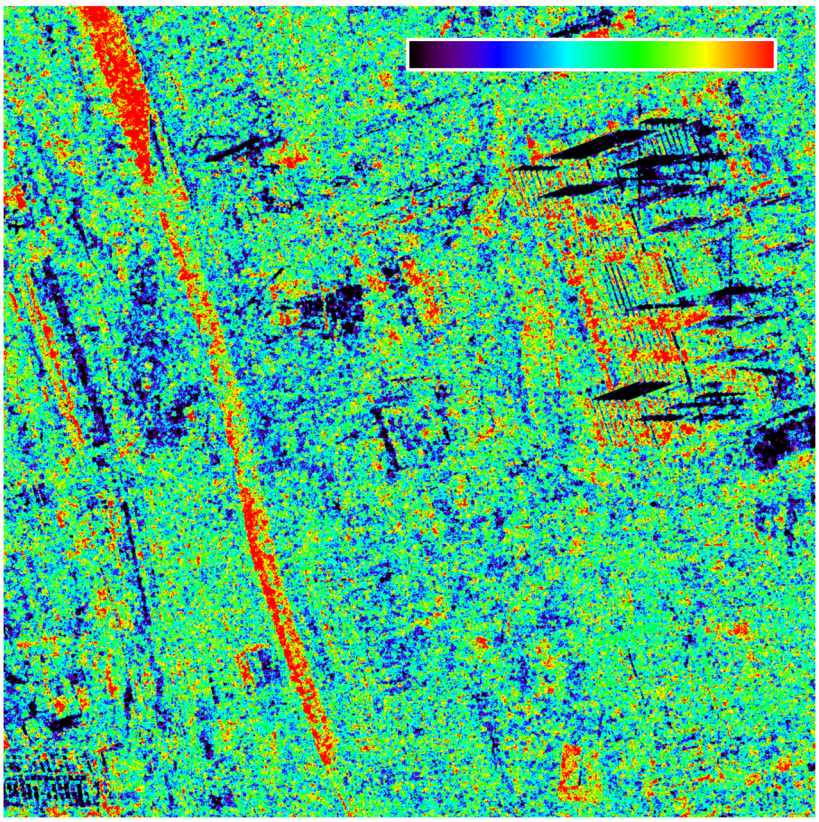


(a)

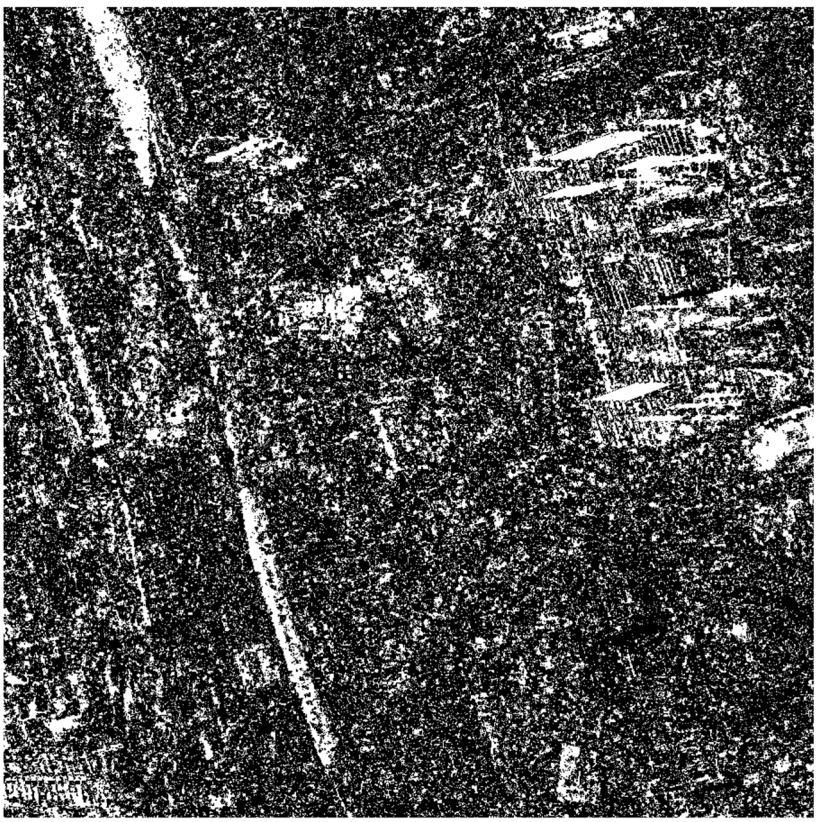


(b)

Figure 18. Original example over patch 1 (Figure 8). (a) Ratio image (potential changes towards black and red). (b) Detected changes (white).

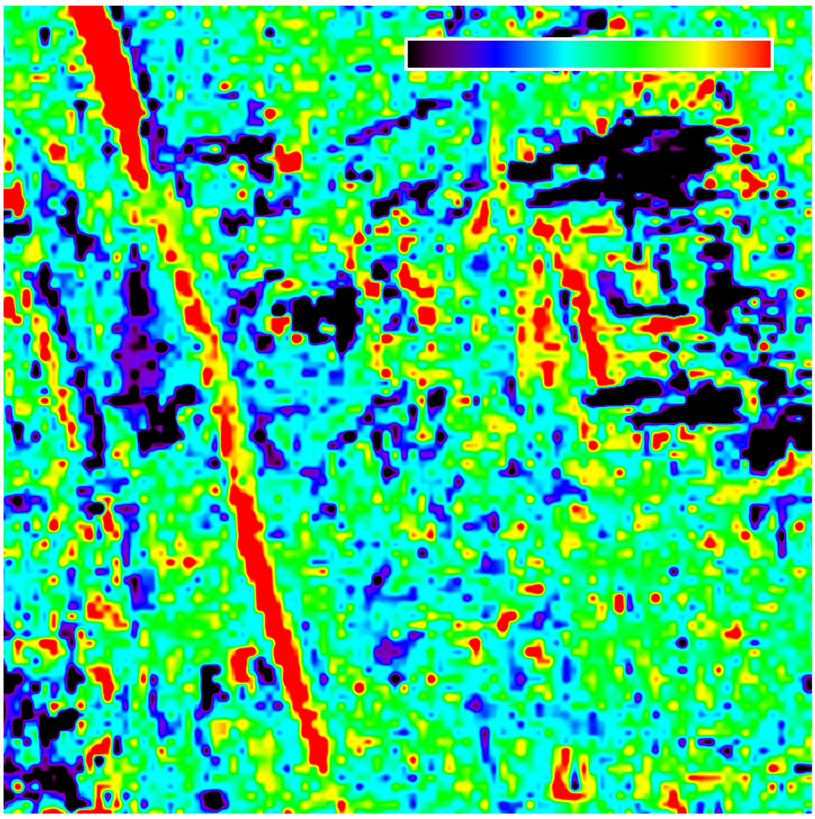


(a)



(b)

Figure 19. Despeckle example over patch 1 (Figure 8). (a) Ratio image (potential changes towards black and red), and (b) Detected changes (white).



(a)

(b)

Figure 20. Multi-looking example over patch 1 (Figure 8). (a) Ratio image (potential changes towards black and red), and (b) Detected changes (white).

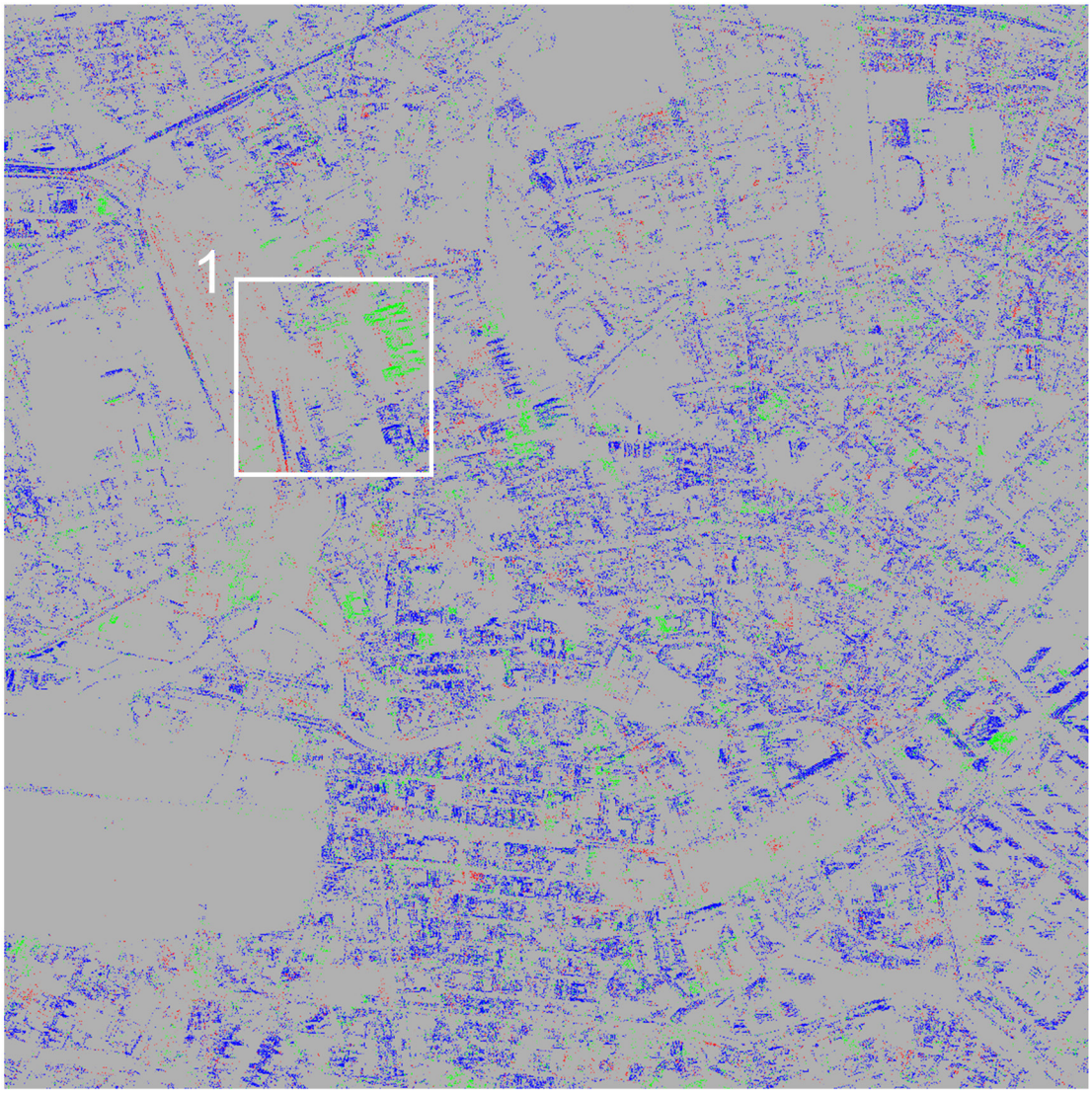


Figure 21. Amplitude-based semi-PS result. Patch 1 is used for in-depth analysis. Steady, disappearing, and emerging structures represented by PS (blue, 49,868/km^2), DBC (red, 5991/km^2), and EBC (green, 9283/km^2) points.

Berardino, P., G. Fornaro, R. Lanari, and E. Sansosti, 2002. A new algorithm for surface deformation monitoring based on small baseline differential SAR interferograms, *IEEE Transactions on Geoscience and Remote Sensing*, 40(11):2375–2383.

Bovenga, F., J. Wasowski, D.O. Nitti, R. Nutricato, and M.T. Chiaradia, 2012. Using COSMO/SkyMed X-band and ENVISAT C-band SAR interferometry for landslides analysis, *Remote Sensing of Environment*, 119:272–285.

Brcic, R., and N. Adam, 2013. Detecting changes in persistent scatterers, *2013 IEEE International Geoscience and Remote Sensing Symposium*, Melbourne, Australia, 117– 120.

Costantini, M., S. Falco, F. Malvarosa, and F. Minati, 2008. A new method for identification and analysis of persistent scatterers in series of SAR images, *Proceedings of the 2008 IEEE International Geoscience and Remote Sensing Symposium*, Boston, Massechuttes.

Crosetto, M., A. Arnaud, J. Duro, E. Biescas, and M. Agudo, 2003. Deformation monitoring using remotely sensed radar interferometric data, *Proceedings of the 11th FIG Symposium*, Santorini, Greece.

Crosetto, M., B. Crippa, and E. Biescas, 2005. Early detection and in-depth analysis of deformation phenomena by radar interferometry, *Engineering Geology*, 79(1-2):81–91.

Crosetto, M., O. Monserrat, M. Cuevas-González, N. Devanthéry, and B. Crippa, 2016. Persistent scatterer interferometry: A review, *ISPRS Journal of Photogrammetry and Remote Sensing*, 115:78–89.

Ferretti, A., C. Colesanti, D. Perissin, C., Prati, and F. Rocca, 2003. Evaluating the effect of the observation time on the distribution of SAR permanent scatterers, *FRINGE 2003*. Frascati, Italy.

Ferretti, A., A. Fumagalli, F. Novali, C. Prati, F. Rocca, and A. Rucci, 2011. A new algorithm for processing interferometric data-stacks: SqueeSAR, *IEEE Transactions on Geoscience and Remote Sensing*, 49(9):3460–3470.

Ferretti, A., C. Prati, and F. Rocca, 2000. Nonlinear subsidence rate estimation using permanent scatterers in differential SAR interferometry, *IEEE Transactions on Geoscience and Remote Sensing*, 38(5):2202–2212.

Ferretti, A., C. Prati, and F. Rocca, 2001. Permanent scatterers in SAR interferometry, *IEEE Transactions on Geoscience and Remote Sensing*, 39(1):8–20.

Gamba, P., 2013. Human settlements: A global challenge for EO data processing and interpretation, *Proceedings of the IEEE*, 101(3):570–581.

Hanssen, R.F., 2001. *Radar interferometry - Data Interpretation and Error Analysis*, Springer, Dordrecht, The Netherlands.

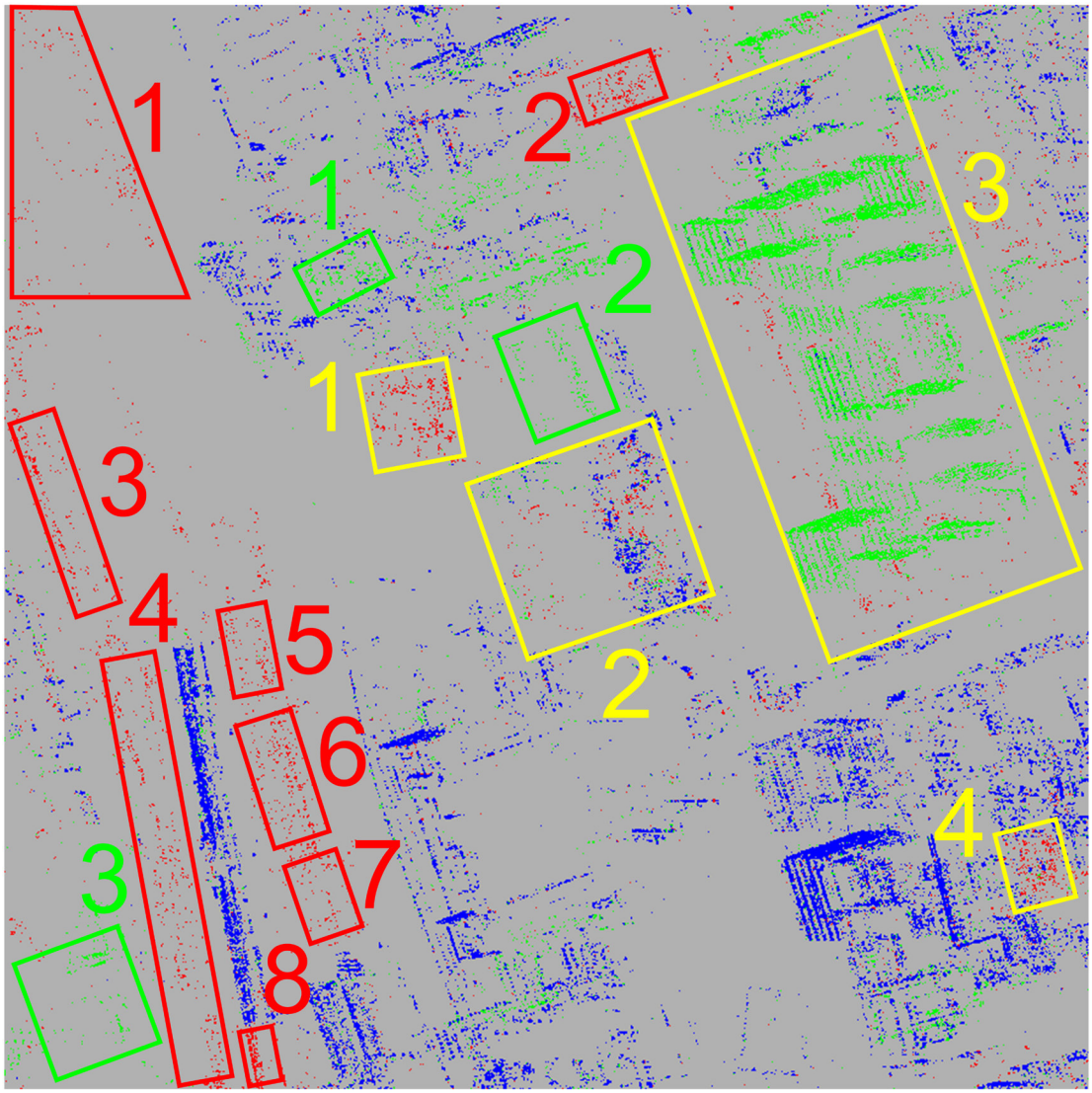


Figure 22. Amplitude-based semi-PS result within patch 1 (Figure 21). Building change: red, disappearance area; green, emergence area; and yellow, hybrid area.

Hooper, A., H. Zebker, P. Segall, and B. Kampes, 2004. A new method for measuring deformation on volcanoes and other natural terrains using InSAR persistent scatterers, *Geophysical Research Letters*, 31(23).

Kampes, B., 2006. *Radar Interferometry - Persistent Scatterer Technique*, Springer, Dordrecht, The Netherlands.

Lanari, R., O. Mora, M. Manunta, J.J. Mallorqui, P. Berardino, and E. Sansosti, 2004. A small-baseline approach for investigating deformations on full-resolution differential SAR interferograms, *IEEE Transactions on Geoscience and Remote Sensing*, 42(7):1377–1386.

Lee, J.-S., 1981. Refined filtering of image noise using local statistics, *Computer Graphics and Image Processing*, 15(4):380–389.

Lyons, S., and D. Sandwell, 2003. Fault creep along the southern san andreas from interferometric synthetic aperture radar, permanent scatterers, and stacking, *Journal of Geophysical Research*, 108(B1).

Marin, C., F. Bovolo, and L. Bruzzone, 2015. Building change detection in multitemporal very high resolution SAR images, *IEEE Transactions on Geoscience and Remote Sensing*, 53(5):2664–2682.

Novali, F.,M. Basilico, A. Ferretti, C. Prati, and F. Rocca, 2004. Advances in permanent scatterer analysis: Semi and temporary PS, *Proceedings of EUSAR 2004*, Berlin, Germany.

Otsu, N., 1979. A threshold selection method from gray-level histograms, *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66.

Rignot, E.J.M., and J.J. van Zyl, 1993. Change detection techniques for ERS-1 SAR data, *IEEE Transactions on Geoscience and Remote Sensing*, 31(4):896–906.

Teunissen, P.J.G., 2000. *Adjustment Theory: An Introduction*, VSSD, Delft, The Netherlands.

Touzi, R., A. Lopes, J. Bruniquel, and P.W. Vachon, 1999. Coherence estimation for SAR imagery, *IEEE Transactions on Geoscience and Remote Sensing*, 37(1):135–149.

Werner, C., U. Wegmuller, T. Strozzi, and A. Wiesmann, 2003. Interferometric point target analysis for deformation mapping, *Proceedings of the* 2003 IEEE International *Geoscience and Remote Sensing Symposium*, Toulouse, France, 4362–4364.

Zebker, H.A., P.A. Rosen, and S. Hensley, 1997. Atmospheric effects in interferometric synthetic aperture radar surface deformation and topographic maps, *Journal of Geophysical Research*, 102(B4):7547–7563.

We'd Like to Give You a Hand

Claim Your FREE Access to Sample Data, Imagery and View/Measure Tools

Go to: www.geomni.net/psm

Geomni

Generate new revenue. Save time. Reduce costs. Make more profit.