DeepStat: Simplifying Deep Learning for Remote Sensing Data in Official Statistics

Keywords: *Official Statistics, Remote Sensing, Deep Learning, IT framework, Big Data*

1 Introduction

In recent years, National Statistical Offices (NSOs) have increasingly searched for information sources alternative and in addition to survey data. Progressively more administrative data sources are used in official statistics. The first statistical products based on big data sources have already been published [1]. A promising big data source are remote sensing data, i.e., satellite or aerial imagery, that can provide a detailed and an integral view of a country's urban areas [2, 3], infrastructure [4], and natural resources [5, 6, 7, 8]. Remote sensing data can be used as a main data source for official statistics, for example regarding the UN sustainable development goals, such as urbanization, poverty, or the production of clean energy [9, 10]. It can also be used as an additional data source, for instance to quality check existing registers, or as a way to extrapolate survey data to geographical areas that are less densely covered.
While already a large body of research has been done in the remote sensing community [11, 12], remote sensing data has not found widespread adoption in official statistics. In a large part, this can be explained by the remote sensing data itself. Unlike more traditional data sources, remote sensing data needs new ways of data processing. First, the amount of data processed is much larger, usually gigabytes or terabytes of data, typically requiring specialized IT hardware. Second, remote sensing data consists of image data that need specific ways of processing to be integrated in official statistics. In this paper, we present the DeepStat framework that aims to simplify this integration and automate as much of the process as possible. This enables statistical researchers to focus on data exploration, information extraction, and the creation of new statistical products based on remote sensing data. The first part in the name **DeepStat** refers to *deep learning*, a method within the domain of machine learning that excels in a variety of domains; the classification and analysis of images being one (i.e., it learns a model that makes predictions on what an image is about, e.g., it is an urban area, grassland, etc.). Deep learning is a fundamental part in the process towards official statistics, hence the name *DeepStat*.The *DeepStat* framework supports the researchers through the whole process starting with data acquisition, data pre-processing, (deep) model training & evaluation, to bridging the gap between model predictions and statistical units.

2 Overview and Overall Goal of the DeepStat framework

The overall goal of the DeepStat framework is to overcome the IT barriers encountered in the use of remote sensing data by creating user-friendly tools that facilitate deep learning experiments on aerial and satellite images in order to produce better and new official statistics using a sound methodology. End users can be supported in creating high quality official statistics while being shielded from the complex technologies and methodologies that lie behind it. Moreover, by standardizing data collection, data

pre-processing, training and evaluation and storing results in a central portal, a lot of complexity, reinventing the wheel, and common mistakes can be avoided. In this sense, the DeepStat framework aims to support at least four phases in the Generic Statistical Business Process Model: (1) Collect, (2) Process, (3) Analyze and (4) Disseminate. As some parts of these phases look different for remote sensing data, we will lay out how the DeepStat framework aims to support these phases in the next subsections.

2.1 Collect

A lot of Remote Sensing data is available as open data, for instance via portals like the Copernicus Open Access Hub[1] or that of the US Geological Survey[2]. To be used in the statistical process, the data need to be collected and possibly extracted from those portals. While a variety of open standards for collecting remote sensing data exists [3], some portals prefer to provide their own APIs. As these APIs are different from other data sources in official statistics, bespoke connections need to be build. Creating a representative sample of remote sensing data is also not straightforward. Although, such a sample can be in principle be based on register labels for a certain statistic, these register labels may not be directly connected to the variety in the underlying data. As such, DeepStat aims to provide multiple ways of creating samples from the data, as for example the creation of a sample on the basis of patterns or similarities in the remote sensing data itself. In this way, the best way of creating representative samples for remote sensing data can be investigated.

2.2 Process

Remote sensing data entails different types of data from different sensors, different parts of the visual spectrum, as well as satellite data and aerial images. Moreover, to use remote sensing data we need to be able to process large quantities of data and make them suitable to be used with machine learning techniques that can detect patterns in the data. Often, remote sensing data also needs to be processed before it can be used, for instance by removing atmospheric disturbances. DeepStat aims to provide support for a variety of remote sensing data, provide commonly used pre-processing techniques out of the box, and automate support as much as possible. More specifically, to process the remote sensing data DeepStat offers various state-of-the-art Deep Learning algorithms [13, 14, 15] that can be trained and evaluated on the data. To train well-performing Deep Learning algorithms so-called hyper-parameters need to be tuned and certain performance metrics need to be optimized [16]. DeepStat offers several hyper-parameter optimization techniques [17, 18] that automatize this process and furthermore offers standard performance metrics out-of-the-box. In addition, all results and models are stored along with their performance metrics, which enables statistical researchers to carry out and compare a range of experiments on various remote data sets.

2.3 Analyze & Disseminate

To use the results of Deep Learning algorithms in official statistics, the resulting model predictions need to be translated into statistical units. This translation is still the subject of current research. Especially, it is not entirely clear how to retrieve

[1] https://scihub.copernicus.eu
[2] https://www.usgs.gov/
[3] https://www.osgeo.org/about/open-standards/

model uncertainty out of deep learning models [19] and prediction probabilities often do not relate to the underlying distribution of classes; i.e. deep learning models are largely uncalibrated [20]. To support research into this areas, DeepStat offers a range of metrics for model uncertainty and model calibration. Furthermore, the Deep Learning algorithms trained need to be evaluated for other geographical regions that have not been encountered during the training process. Several methods of model evaluation have been suggested, of which the cross-region evaluation gives the best indication of how well models generalize [9, 21]. To this cause, DeepStat simplifies model evaluation on other geographical regions and datasets. Finally, to disseminate results, DeepStat offers a way to visualize results geographically, create interactive plots of model results, and relate them to other register data available.

3 Requirements and Implementation

DeepStat is implemented as open source software and builds upon existing open source frameworks for remote sensing. It integrates several local and international open data sources, the Dutch PDOK geographical portal, Copernicus Open Access Hub[4], and the US Geological Survey[5]. By adhering to open geographical standards like those provided by the Open Source Geospatial Foundation [6] DeepStat can integrate other third-party geographical data sources more easily. In addition, by using generic open software standards interoperability with other geographical systems can be provided and maintained. DeepStat will be implemented in a client-server infrastructure, where researchers can connect to a central portal that stores all their data and experiments. By using a standard like for example OpenAPI[7] to specify the client server interaction, the connection to a variety of programming languages can be facilitated and automated. Moreover, integration with other data processing tools will be simplified. To evaluate whether the DeepStat framework is fit for the tasks at hand and provide proper support, it will be tested with end-users in the statistical departments on a variety of geo-spatial use-cases related to the Sustainable Development Goals. This is ensured by an iterative implementation process that puts an evaluation with end-users at the end of each implementation cycle.

4 Conclusions

This paper presented the DeepStat framework, a framework that makes remote sensing data more accessible to statistical researchers. DeepStat aims to shield the statistical researchers as much as possible from the technical details and complexity of analysing remote sensing data using deep learning methods. It helps researchers create more reproducible research by providing a central storage for all data, results, and evaluations. By offering standard tools it also helps researchers with the methodological issues that might be encountered in the use of remote sensing data. In this way, DeepStat's main objective is a more widespread use of remote sensing data as an additional or primary source of information for official statistics.

[4] https://scihub.copernicus.eu
[5] https://www.usgs.gov/
[6] https://www.osgeo.org/about/open-standards/
[7] https://github.com/OAI/OpenAPI-Specification

References

[1] M.J.H. Puts, P.J.H. Daas, M. Tennekes, and C. de Blois. Using huge amounts of roadsensor data for official statistics. *AIMSMathematics*, 4:12–25, 2019.

[2] Renaud Mathieu, Claire Freeman, and Jagannath Aryal. Mapping private gardens in urban areas using object-oriented techniques and very high-resolution satellite imagery. *Landscape and Urban Planning*, 81(3):179 – 192, 2007. ISSN 0169-2046. DOI: https://doi.org/10.1016/j.landurbplan.2006.11.009. URL http://www.sciencedirect.com/science/article/pii/S0169204606002684.

[3] Maryam Malmir, Mir Masoud Kheirkhah Zarkesh, Seyed Masoud Monavari, Seyed Ali Jozi, and Esmail Sharifi. Urban development change detection based on multi-temporal satellite images as a fast tracking approach—a case study of ahwaz county, southwestern iran. *Environmental monitoring and assessment*, 187 (3):108, 2015.

[4] Donna Haverkamp. Extracting straight road structure in urban environments using ikonos satellite imagery. *Optical Engineering*, 41(9):2107–2110, 2002.

[5] MX Ortega, JD Bermudez, PN Happ, A Gomes, and RQ Feitosa. Evaluation of deep learning techniques for deforestation detection in the amazon forest. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, 4, 2019.

[6] Jorge Caballero Espejo, Max Messinger, Francisco Román-Dañobeytia, Cesar Ascorra, Luis Fernandez, and Miles Silman. Deforestation and forest degradation due to gold mining in the peruvian amazon: A 34-year perspective. *Remote Sensing*, 10(12):1903, Nov 2018. ISSN 2072-4292. DOI: 10.3390/rs10121903. URL http://dx.doi.org/10.3390/rs10121903.

[7] Matteo G. Ziliani, Stephen D. Parkes, Ibrahim Hoteit, and Matthew F. McCabe. Intra-season crop height variability at commercial farm scales using a fixed-wing uav. *Remote Sensing*, 10(12), 2018. ISSN 2072-4292. DOI: 10.3390/rs10122007. URL https://www.mdpi.com/2072-4292/10/12/2007.

[8] Lillian Kay Petersen. Real-time prediction of crop yields from modis relative vegetation health: A continent-wide analysis of africa. *Remote Sensing*, 10(11), 2018. ISSN 2072-4292. DOI: 10.3390/rs10111726. URL https://www.mdpi.com/2072-4292/10/11/1726.

[9] Rui Wang, Joseph Camilo, Leslie M. Collins, Kyle Bradbury, and Jordan M. Malof. The poor generalization of deep convolutional networks to aerial imagery from new geographic locations: an empirical study with solar array detection. In *2017 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–8, Washington, DC, USA, October 2017. IEEE. ISBN 978-1-5386-1235-4. DOI: 10.1109/AIPR.2017.8457960. URL https://ieeexplore.ieee.org/document/8457960/.

[10] Jiafan Yu, Zhecheng Wang, Arun Majumdar, and Ram Rajagopal. DeepSolar: A Machine Learning Framework to Efficiently Construct a Solar Deployment Database in the United States. *Joule*, 2(12):2605–2617, December 2018. ISSN 2542-4785, 2542-4351. DOI: 10.1016/j.joule.2018.11.021. URL https://www.cell.com/joule/abstract/S2542-4351(18)30570-1.

[11] Michael A. Wulder, Thomas R. Loveland, David P. Roy, Christopher J. Crawford, Jeffrey G. Masek, Curtis E. Woodcock, Richard G. Allen, Martha C. Anderson, Alan S. Belward, Warren B. Cohen, John Dwyer, Angela Erb, Feng Gao, Patrick Griffiths, Dennis Helder, Txomin Hermosilla, James D. Hipple, Patrick Hostert, M. Joseph Hughes, Justin Huntington, David M. Johnson, Robert Kennedy, Ayse Kilic, Zhan Li, Leo Lymburner, Joel McCorkel, Nima Pahlevan, Theodore A. Scambos, Crystal Schaaf, John R. Schott, Yongwei Sheng, James Storey, Eric Vermote, James Vogelmann, Joanne C. White, Randolph H. Wynne, and Zhe Zhu. Current status of Landsat program, science, and applications. *Remote Sensing of Environment*, 225(February):127–147, 2019. ISSN 00344257. DOI: 10.1016/j.rse.2019.02.015.

[12] Pratistha Kansakar and Faisal Hossain. A review of applications of satellite earth observation data for global societal benefit and stewardship of planet earth. *Space Policy*, 36:46 – 54, 2016. ISSN 0265-9646. DOI: https://doi.org/10.1016/j.spacepol.2016.05.005. URL http://www.sciencedirect.com/science/article/pii/S0265964616300133.

[13] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. pages 1–14, 2014. ISSN 09505849. DOI: 10.1016/j.infsof.2008.09.005. URL http://arxiv.org/abs/1409.1556.

[14] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, CVPR, pages 770–778. IEEE, 2016.

[16] Amalia Luque, Alejandro Carrasco, Alejandro Martín, and Ana de las Heras. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, 91:216–231, July 2019. ISSN 0031-3203. DOI: 10.1016/j.patcog.2019.02.023. URL http://www.sciencedirect.com/science/article/pii/S0031320319300950.

[17] Krzysztof Cybulski. Declair. https://gitlab.com/k-cybulski/declair, 2020.

[18] Tom O'Malley, Elie Bursztein, James Long, François Chollet, Haifeng Jin, Luca Invernizzi, et al. Keras Tuner. https://github.com/keras-team/keras-tuner, 2019.

[19] Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? page 11, 2017.

[20] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, volume 70 of *ICML'17*, pages 1321–1330, Sydney, NSW, Australia, June 2017. JMLR.org. URL http://arxiv.org/abs/1706.04599. arXiv: 1706.04599.

[21] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can Semantic Labeling Methods Generalize to Any City? The Inria Aerial Image Labeling Benchmark. In *IEEE International Symposium on Geoscience and Remote Sensing (IGARSS)*, Fort Worth, United States, July 2017. URL https://hal.inria.fr/hal-01468452.