This is an excerpt from the thesis "*Multi-Output Regression: On the Impact of Individual Model Parameters for Built-Up Height and Density Prediction*".

Please contact Elisabeth Brzoska for a full version of the thesis.

Extended works on this topic are documented in: Geiß, C., Brzoska, E., Aravena Pelizari, P., Lautenbach, S., and Taubenböck, H. (2022): Multi-target Regressor Chains with Repetitive Permutation Scheme for Characterization of Built Environments with Remote Sensing. International Journal of Applied Earth Observation and Geoinformation, 106, 102657.

Multi-Output Regression: On the Impact of Individual Model Parameters for Built-Up Height and Density Prediction

Heidelberg 2020

Abstract

Urbanisation is an ongoing process and will gain importance in the future. It comes with multiple challenges as inhabitants are dependent on water and energy supply, a functioning street network and health care system — all these require a deliberate management. However, this is not an easy demand as administrative areas can cover several thousands of square kilometers. Therewith, remote sensing methods constitute a reliable source to observe large areas as cities. To observe the growth of cities, variables as built-up height and built-up density have emerged as reliable attributes that characterize well the urban morphology. They can be obtained by integrating remote sensing data from optical and other sensors such as synthetic aperture radar (SAR). The application of machine learning algorithms makes it feasible to interpret the large amount of data generated in remote sensing.

This study focuses on the optimization of machine learning algorithms for predicting built-up height and built-up density in four German major cities based on remote sensing data, by integrating so-called multi-output regression (MOR) methods. Instead of processing and predicting each target variable independently, MOR methods incorporate all target variables into one process which, in the best case, increases the accuracy of predictions. Recent literature highlights the benefit of exploiting possible correlations between target variables. In this work, four methods are applied and modified according to state-of-the-art models: multi-target stacking (MTS), multi-target regressor chains (MTRC), multi-target regressor chains without repetitive permutation (MTRC-nrp) and single-target stacking (STS). Each method is used with four different regression models, namely random forest (RF), Gaussian process (GP), support vector regression (SVR) and neural networks (NN). Additionally, the impact of different stacking options as well as the impact of the feature space is evaluated.

The extensive and systematic evaluation of the aforementioned parameters provides several insights. It shows, that all models (MTS, MTRC, MTRC-nrp, STS) outperform models that do not use multi-target stacking or chaining or single-target stacking. Furthermore, it shows that MOR models behave differently depending on which regression model is used for the prediction. Finally, it gives recommendations on which MOR methods and which additional parameters are suitable for particular use cases similar to those evaluated in this study and discusses possibilities for future research.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Our population is growing rapidly and so are the challenges accompanying this process. Challenges of growing cities and the demand of living space, of growing food demand, of energy and water supply, of security and health. As the United Nations prospect, a number of 9.8 billion people will inhabit the world by the year 2050 of which 68% will live in urban areas. [United Nations and Affairs, 2019] This makes the necessity of a good organisation and a good management of societies and communities on a local level to a task, we have to start tackling today to be prepared for changes that will confront us in the future.

Growing cities imply a growing floor space covered by buildings and other man-made structures and thus the necessity of suitable methods to investigate an area which already nowadays can measure several thousand square kilometers (e.g. Tokyo: 2.2 thousand km^2, Rome 1.3 thousand km^2, Hong Kong: 1.1 thousand km^2). It is evident that methods which overlook large areas from above outperform those working in-situ, as they manage to retrieve information much faster and consistently over time and space – such as remote sensing methods do. In remote sensing, satellite imagery is retrieved at daily intervals from numerous satellite missions such as Sentinel, Landsat or TanDEM-X. They provide freely available data from multiple sensors, for example optical or radar, in a spatial resolution of up to 10 m. [Attema et al., 2007], [Gascon et al., 2014] This makes remote sensing a popular and convenient method for the inspection of urbanisation processes.

From an aerial perspective, an urban area is basically characterized by its physical structures, as the network of streets, the distribution of green areas or the shape and height of buildings and building complexes. Especially the latter, the attributes of built-up area like built-up height and built-up density, constitute a basis in urban morphology research. They give a strong indication on where population in urban areas is concentrated and which areas are only sparsely populated. For organisational challenges, these information are crucial. Therewith, they also gain importance in remote sensing research and studies aim to develop methods which derive built-up attributes from remote sensing data (e.g. [Brunner et al., 2009], [Guida et al., 2010], [Kim et al., 2007], [Kajimoto and Susaki, 2013]). It is a common approach to integrate remote sensing data of various sensors to increase the level of information. The integration of data derived by spaceborne Synthetic Aperture Radar (SAR), for example, can augment common satellite imagery by information on height of recorded objects. Conventional satellite imagery, which is derived from missions such as Sentinel or Landsat, focus on optical information. Their sensors record reflected solar or thermal radiation in multiple bands where each focuses on a specific frequency range. A common way to derive information from these data is to combine the actual values recorded by particular bands of interest to so-called features and map them according to their geographical appearance. With the help of statistical learning methods, such as conventional machine learning algorithms, it is possible to identify correlations between calculated features and the actual target variable, such as built-up attributes.

Machine learning methods constitute a powerful set of algorithms, which are capable of learning and improving from the experience of their performance on particular tasks. [Mitchell et al., 1997] Based on their experience, they can make predictions on the expected behaviour in contexts similar to their learning environment. Therewith, these methods can strongly support the investigation of data which is too complex or too large to be investigated manually. [Mitchell, 2006] Applying machine learning methods in the area of remote sensing boosts the possibilities of aerial analysis remarkably. They make it feasible to process a large amount of data and to generate information in areas, in which appropriate data might not be available or accessible. The combination of both, machine learning and remote sensing, gives scientists and political stakeholders the opportunity to observe, investigate and regulate processes on a local, regional and global level and on a temporal and spatial consistent basis. Processes, such as the ongoing process of urbanisation.

This work aims to contribute to existing methodological approaches of predicting built-up height and built-up density in remote sensing with the application of machine learning methods. It will modify existing approaches and conduct a systematic and extensive evaluation on the impact of particular experimental parameters.
As two variables, built-up height and density, constitute the target of this study, the focus will lay on the application of so-called multi-output regression (MOR) methods. A multi-output regression, or multi-target regression, as it is also called in literature, describes a method which is constructed to predict multiple target variables. In the best case, the prediction of each single target variable can benefit from information gained trough the involvement of the other target variables in the regression process. Several authors have already applied and tested MOR on various datasets.
Numerous studies exist, which apply existing MOR methods on their explicit research problem of interest or on several datasets, and evaluate and compare the performance of tested MOR methods (e.g. [Tuia et al., 2011], [Stojanova et al., 2010], [Kocev et al., 2009], [Melki et al., 2017], [Segal and Xiao, 2011]). [Tuia et al., 2011], for example, analyze the prediction of the biophysical parameters leaf chlorophyll content, leaf area index and fractional vegetation cover by adapting a support vector regression algorithm to predict multiple target variables (M-SVR). They compare it to a conventional support vector regression (SVR), where each parameter is predicted separately and found that the M-SVR

approach outperforms the conventional SVR when target variables are correlated. [Santana et al., 2017] adapts an existing multi-target stacking (MTS) approach to predict air ticket prices. They claim, that their adapted MTS approach outperforms conventional methods such as single-target regression, in which each variable is predicted independently. MOR methods are also applied explicitly in remote sensing. [Stojanova et al., 2010], for example, apply state-of-the-art MOR algorithms on the estimation of vegetation height and canopy cover. Their findings show a benefit of ensemble MOR algorithms over single-target models.

On the other hand, studies exist which focus on a theoretical perspective on MOR. [Borchani et al., 2015], [Spyromitros-Xioufis et al., 2012] and [Waegeman et al., 2019] for example focus on the methodological part of MOR. All three give a detailed theoretical introduction on state of the art MOR methods. [Spyromitros-Xioufis et al., 2012] focus on the adaptation of multilabel classification methods on regression problems and introduce suitable multi-output stacking and chaining methods. They outline the importance of including cross-validation in the process of test and training set splitting and emphasize the benefit of applying the presented MOR methods over conventional single-target methods. While [Borchani et al., 2015] mainly focus on the presentation of MOR methods, datasets and an overview on advantages and disadvantages of particular methods, [Waegeman et al., 2019] goes beyond that and postulates the need of putting more effort in the investigation of MOR to increase deeper understanding.

The here presented work is conceptualized to address this demand. It concentrates on the particular problem of predicting built-up height and built-up density based on remote sensing imagery and applies suitable MOR methods. The chosen methods are then systematically tested and each relevant parameter constituting the experiment configuration is evaluated. Subsequently, this work should be an attempt to the investigation of the actual behaviour of multi-output regression methods.

This work is structured as follows: Chapter 2 will give a theoretical introduction in the applied methods. It will first outline the functioning of multi-output regression methods (section 2.1) and continues with a theoretical description of the machine learning methods used in this study (section 2.2). Chapter 3 first describes the input data (section3.1) and then focuses on the description of all parameters that constitute the conducted experiments (section 3.2). It gives a clear definition on how MOR methods are applied and modified in the context of this study (subsection 3.2.1 - 3.2.3) and finally outlines the experiments that have been used for evaluation (section 3.2.4). Subsequently, chapter 4 will combine the presentation of all results as well as the discussion on the presented findings. Lastly, chapter 5 will conclude the work and give an outlook on suggested follow up research.

This is an excerpt from the thesis "*Multi-Output Regression: On the Impact of Individual Model Parameters for Built-Up Height and Density Prediction*".

Please contact Elisabeth Brzoska for a full version of the thesis.

Extended works on this topic are documented in: Geiß, C., Brzoska, E., Aravena Pelizari, P., Lautenbach, S., and Taubenböck, H. (2022): Multi-target Regressor Chains with Repetitive Permutation Scheme for Characterization of Built Environments with Remote Sensing. International Journal of Applied Earth Observation and Geoinformation, 106, 102657.

Bibliography

[Attema et al., 2007] Attema, E., Bargellini, P., Edwards, P., Levrini, G., Lokas, S., Moeller, L., Rosich-Tell, B., Secchi, P., Torres, R., Davidson, M., et al. (2007). Sentinel-1-the radar mission for gmes operational land and sea services. *ESA bulletin*, 131:10–17.

[Awad and Khanna, 2015] Awad, M. and Khanna, R. (2015). Support vector regression. In *Efficient Learning Machines*, pages 67–80. Springer.

[Bergmeir and Benítez, 2012] Bergmeir, C. and Benítez, J. M. (2012). Neural networks in R using the stuttgart neural network simulator: RSNNS. *Journal of Statistical Software*, 46(7):1–26.

[Bernard et al., 2009] Bernard, S., Heutte, L., and Adam, S. (2009). Influence of hyper-parameters on random forest accuracy. In *International Workshop on Multiple Classifier Systems*, pages 171–180. Springer.

[Bezirksregierung Köln, 2014] Bezirksregierung Köln (2014). Datenformatbeschreibung 3D-Gebäudemodell LoD1 Deutschland. Hauskoordinaten und Hausumringe des deutschen Liegenschaftskatasters, Version 1.0 Stand: 11.07.2014

[Borchani et al., 2015] Borchani, H., Varando, G., Bielza, C., and Larrañaga, P. (2015). A survey on multi-output regression. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(5):216–233.

[Brunner et al., 2009] Brunner, D., Lemoine, G., Bruzzone, L., and Greidanus, H. (2009). Building height retrieval from vhr sar imagery based on an iterative simulation and matching technique. *IEEE Transactions on Geoscience and Remote Sensing*, 48(3):1487–1504.

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

[Dembczyński et al., 2012] Dembczyński, K., Waegeman, W., Cheng, W., and Hüllermeier, E. (2012). On label dependence and loss minimization in multi-label classification. *Machine Learning*, 88(1-2):5–45.

[Dimitriadis et al., 2018] Dimitriadis, S. I., Liparas, D., Initiative, A. D. N., et al. (2018). How random is the random forest? random forest algorithm on the service of structural imaging biomarkers for alzheimer's disease: from alzheimer's disease neuroimaging initiative (adni) database. *Neural regeneration research*, 13(6):962.

[Esch et al., 2017] Esch, T., Heldens, W., Hirner, A., Keil, M., Marconcini, M., Roth, A., Zeidler, J., Dech, S., and Strano, E. (2017). Breaking new ground in mapping human settlements from space–the global urban footprint. *ISPRS Journal of Photogrammetry and Remote Sensing*, 134:30–42.

[Friedman et al., 2001] Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

[Gascon et al., 2014] Gascon, F., Cadau, E., Colin, O., Hoersch, B., Isola, C., Fernández, B. L., and Martimort, P. (2014). Copernicus sentinel-2 mission: products, algorithms and cal/val. In *Earth observing systems XIX*, volume 9218, page 92181E. International Society for Optics and Photonics.

[Geiß et al., 2018] Geiß, C., Schrade, H., and Taubenböck, H. (24-26 September 2018). Boosted machine learning ensemble regression with decision fusion strategy for mapping built-up height and built-up density with openstreetmap data and sentinel-2 imagery. 5th EARSeL Joint Workshop 'Urban Remote Sensing: Challenges and Solutions'.

[Géron, 2017] Géron, A. (2017). *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems.* O'Reilly Media, Inc.

[Ghafari et al., 2012] Ghafari, E., Bandarabadi, M., Costa, H., and Júlio, E. (2012). Design of uhpc using artificial neural networks. In *Brittle Matrix Composites 10*, pages 61–69. Elsevier.

[Ghojogh and Crowley, 2019] Ghojogh, B. and Crowley, M. (2019). The theory behind overfitting, cross validation, regularization, bagging, and boosting: tutorial. *arXiv preprint arXiv:1905.12787*.

[Görtler et al., 2019] Görtler, J., Kehlbeck, R., and Deussen, O. (2019). A visual exploration of gaussian processes. *Distill*. https://distill.pub/2019/visual-exploration-gaussian-processes.

[Guida et al., 2010] Guida, R., Iodice, A., and Riccio, D. (2010). Height retrieval of isolated buildings from single high-resolution sar images. *IEEE Transactions on Geoscience and Remote Sensing*, 48(7):2967–2979.

[Gunn et al., 1998] Gunn, S. R. et al. (1998). Support vector machines for classification and regression. *ISIS technical report*, 14(1):5–16.

[Hall, 1999] Hall, M. A. (1999). Correlation-based feature selection for machine learning.

[Kajimoto and Susaki, 2013] Kajimoto, M. and Susaki, J. (2013). Urban density estimation from polarimetric sar images based on a poa correction method. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6(3):1418–1429.

[Karatzoglou et al., 2004] Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004). kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9):1–20.

[Kim et al., 2007] Kim, T., Javzandulam, T., and Lee, T.-Y. (2007). Semiautomatic reconstruction of building height and footprints from single satellite images. In *2007 IEEE International Geoscience and Remote Sensing Symposium*, pages 4737–4740. IEEE.

[Kocev et al., 2009] Kocev, D., Džeroski, S., White, M. D., Newell, G. R., and Griffioen, P. (2009). Using single-and multi-target regression trees and ensembles to model a compound index of vegetation condition. *Ecological Modelling*, 220(8):1159–1168.

[Lever et al., 2016] Lever, J., Krzywinski, M., and Altman, N. (2016). Points of significance: model selection and overfitting.

[Liaw and Wiener, 2002] Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.

[Mateo et al., 2015] Mateo, F., Muñoz-Marí, J., Laparra, V., Verrelst, J., and Camps-Valls, G. (2015). Learning structures in earth observation data with gaussian processes. In *International Workshop on Advanced Analysis and Learning on Temporal Data*, pages 78–94. Springer.

[Melki et al., 2017] Melki, G., Cano, A., Kecman, V., and Ventura, S. (2017). Multi-target support vector regression via correlation regressor chains. *Information Sciences*, 415:53–69.

[Meyer et al., 2019] Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2019). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.7-3.

[Mitchell, 2006] Mitchell, T. M. (2006). *The discipline of machine learning*, volume 9. Carnegie Mellon University, School of Computer Science, Machine Learning .

[Mitchell et al., 1997] Mitchell, T. M. et al. (1997). Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, 45(37):870–877.

[Romanski and Kotthoff, 2018] Romanski, P. and Kotthoff, L. (2018). *FSelector: Selecting Attributes*. R package version 0.31.

[Santana et al., 2017] Santana, E., Mastelini, S., et al. (2017). Deep regressor stacking for air ticket prices prediction. In *Anais do XIII Simpósio Brasileiro de Sistemas de Informação*, pages 25–31. SBC.

[Segal and Xiao, 2011] Segal, M. and Xiao, Y. (2011). Multivariate random forests. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):80–87.

[Segal, 2004] Segal, M. R. (2004). Machine learning benchmarks and random forest regression.

[Smola and Schölkopf, 2004] Smola, A. J. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222.

[Spyromitros-Xioufis et al., 2012] Spyromitros-Xioufis, E., Tsoumakas, G., Groves, W., and Vlahavas, I. (2012). Multi-label classification methods for multi-target regression. *arXiv preprint arXiv:1211.6581*, pages 1159–1168.

[Statistisches Bundesamt, 2019] Statistisches Bundesamt Die größten Städte in Deutschland nach Einwohnerzahl zum 31. Dezember 2018

[Stojanova et al., 2010] Stojanova, D., Panov, P., Gjorgjioski, V., Kobler, A., and Džeroski, S. (2010). Estimating vegetation height and canopy cover from remotely sensed data with machine learning. *Ecological Informatics*, 5(4):256–266.

[Tuia et al., 2011] Tuia, D., Verrelst, J., Alonso, L., Pérez-Cruz, F., and Camps-Valls, G. (2011). Multioutput support vector regression for remote sensing biophysical parameter estimation. *IEEE Geoscience and Remote Sensing Letters*, 8(4):804–808.

[United Nations and Affairs, 2019] United Nations, D. o. E. and Affairs, S. (2019). World urbanization prospects: The 2018 revision.

[Van der Aalst et al., 2010] Van der Aalst, W. M., Rubin, V., Verbeek, H., van Dongen, B. F., Kindler, E., and Günther, C. W. (2010). Process mining: a two-step approach to balance between underfitting and overfitting. *Software & Systems Modeling*, 9(1):87.

[Waegeman et al., 2019] Waegeman, W., Dembczyński, K., and Hüllermeier, E. (2019). Multi-target prediction: a unifying view on problems and methods. *Data Mining and Knowledge Discovery*, 33(2):293–324.