First draft research proposal

Niels van Opstal 15-9-2021

Supervisor: Sharon Ong

Project definition

This thesis will try to find out to what extent cellular automata (CA) based land use change (LUC) models can be used to predict densification in the Netherlands. It will specifically look at the way deep learning can improve predictive capabilities. Furthermore this thesis will identify to what extent clustering can help with the heterogenetic nature of different cities and areas. Lastly, this this thesis will try to identify to what extent different bagging methods can help deal with the inherent class imbalance problem in LUC modeling. It will try to answer these questions by analyzing CBS and BAG data.

Motivation

In the next ten years, 900.000 new houses are needed to be built in the Netherlands to face the housing deficit (Ollongren, 2021). The Netherlands prefers new houses to be built within existing urban area (Kenniscentrum InfoMil, n.d.). Dembski et al. (2020) define the net increase of housing units in a given area as densification. They furthermore explain that densification is triggered by reurbanisation, environmental concerns, landscape protection and agricultural land preservation. Modeling urban densification can provide "… researchers working in land change science with important information into urban densification process modeling" as well as enable planners "to make informed decisions to promote planning objectives, which could benefit sustainable urbanization" (Wang et al., 2019, p. 18). Thus this study has a large societal relevance by hopefully giving urban planners and scientists extra handles to tackle the current housing crisis in the Netherlands.

There have been several studies that use Land Use Change (LUC) models to predict changes in land use (e.g., Shafizadeh-Moghadam et al., 2017; Xing et al., 2020; Zhai et al., 2020) and Cellular Automata (CA) based models are very popular (Wang et al., 2019). However, most LUC models are only used to model changes in land type or use such as industrial to residential or green area to urban. Wang et al. (2019) point out that not much attention has been paid to modeling changes in urban density and then shows that LUC can be used to model urban density change. This study will build upon the research of Wang et al. and will try to see to what extent improvements to LUC modeling can also be used in the modeling of urban densification. Deep learning has been showed to work very well for LUC modeling (for example, (Xing et al., 2020)) which leads to the following research question:

RQ: To what extent can deep learning improve the predictive capabilities of a CA-based LUC model for modelling urban densification?

Xing et al. (2020) shows an interesting model where a CNN is used to capture latent spatial features in a CA model and an RNN to capture the temporal features of urban densification. This leads to the following sub questions:

SQ1: To what extent can a CNN be used to optimize a CA based LUC model for modelling urban densification?

SQ2: To what extent can a RNN be used to optimize a CA based LUC model for modelling urban densification?

Then there remain two more problems in LUC modelling that remain largely unanswered in literature, although there have been made some attempts to solve these problems. Firstly, as Omrani et al., (2017) point out, future work should try to address the class imbalance problem inherent in LUC modelling. Xing et al. (2020), for example, addresses this problem by introducing a bagging solution in the form of a random forest (RF) classifier. This leads to the following sub question:

SQ3: To what extent does a bootstrapping method help in addressing the class imbalance problem inherent in modelling urban densification?

Secondly, urban areas are heterogeneous (Cadenasso et al., 2007). There has however not been paid much attention to this problem. One example that does deal with the heterogeneity is Omrani et al. (2019) where the authors perform clustering before training the models and in that way making the data more homogeneous. This leads to the following sub question:

SQ4: To what extent does clustering before training the model lead to improved predictive capabilities of a LUC model for urban densification?

In conclusion, this paper has a big scientific relevance by firstly looking at LUC models for urban densification rather than land change. Secondly by integrating several separate ideas that have been explored on their own for increasing the performance of LUC models.

Background

A lot of the current LUC models are CA-based. Most of the LUC modeling is about land status change such as forest to urban. However, as Wang et al. (2019) showed, the LUC models can be used to also model densification processes. Current LUC models differ mostly in how they calibrate the CA rules. Here is a quick overview of some interesting and recent models and how they each tackled issues with the CA model. Firstly, Shafizadeh-Moghadam et al. (2017) compared statistical, machine learning and tree based models integrated with CA. Here, the models were used to create a suitability map which indicated how suitable each cell is to change. Afterwards, the CA iterates over the maps to calculate for each cell the transition potential based on its neighbors. So the CA operates

as a spatial filter. Afterwards, the cells were ranked based on their potential and changed for the next iteration in order until the number of max changes was reached. They found that all models had an acceptable level of accuracy and they stress the importance of parameter tuning and basing model choice on each individual situation such as is explainability required or not. Shafizadeh-Moghadam (2019) found that using different kind of models together by median ensemble forecasting slightly improved accuracy.

Zhai et al. (2020) used a CNN to optimize the transition probabilities for each cell. They used a CNN specifically since the transition suitability is influenced by a cell's neighbors and a CNN can capture high level-features of a cell's neighborhood. Later on they use the CA again as a spatial filter in same way Shafidez-Moghadam (2017) did. Xing et al. (2020) also uses a CNN but in a different way and also takes the deep-learning component of the CA model a step further. They used the CNN to extract latent spatial features from a cell's neighbors and concatenates them with other variables and uses a random forest (RF) to calculate the transition suitability which automatically includes the neighborhood influences. Here the CNN acts as the neighborhood component of the CA where Zhai et al. uses the CNN as a way to calculate the transition suitability. After that, Xing et al. uses a LSTM to capture the temporal dependencies. The transition suitability predicted by the LSTM is finally added together with constraint factors as well as stochastic factors. It showed that the usage of CNN plus LSTM was very good at modeling LUC.

Wang et al. (2019) used the Land Transformation Model to model urban densification. They used a neural network to learn the transition rules and then model urban density changes in different time steps. They do however not release any specifics about the neural network they used.

Xing et al. (2020) use a RF as a bootstrapping mechanism to deal with the class imbalance inherent in LUC modeling, i.e., most land doesn't change in between time steps. The RF predicts the chance of whether or not a cell will change. RF have the advantage of being bagging ensembles of decision trees that directly implement bootstrap sampling and aggregating rather than having to wrap them into a bagging scheme. Another option that has the same advantage is extreme random forest. (Du et al., 2018)

Lastly, Omrani et al. (2019) make an interesting contribution by clustering the input data and training a model per cluster. By clustering the data they make sure that each cluster is more homogenous (having the same distributions). They found that clustering the data before training outperforms a model trained on not-clustered data as well as that clustering the data makes the size of the data more manageable. This does make sense as Cadenasso et al. (2007) points out that urban areas are heterogeneous.

Data

The data that will be used to generate different densities of residential spaces in a grid comes from the Basisregistraties adressen en gebouwen (BAG) (translation: basis registration addresses and buildings). Dutch municipalities have to deliver the data to BAGLV and the cadaster manages the data. The dataset contains historical data on each building and the different "objects" at each buildings such as a companies or residential spaces in the Netherlands. The data came in two XML files (buildings and objects in buildings) together having a size of 60GB but this became less when the data was processed to only contain the necessary data. Data for the driver variables come from the Central Bureau of Statics. The size and sources of the data can be seen in table **XXXX.** The data will be created in the form of a 100 x 100 m grid per year

Data source	**what**	**size**	**Scale of data**	**Time scale**
BAG	Residential spaces	8 gb	Individual buildings	complete
CBS	Socio-economic data, social security, demographic, nearness of amenities	200 mb per year	100 x 100 m grid	2014 – 2018

Algorithms and evaluations

This study will somewhat follow the methodology of Xing et al. (2020). First, several LUC-CA models will be trained (SVM-CA, RF-CA and LR-CA) which will act as a baseline for comparing predictive power of a LUC model for densification. Then to answer the research question a CNN-CA, RNN-CA, CNN-RNN-CA will be trained. Followed by the addition of clustering and/or a RF as a bootstrapping algorithm. To evaluate the performance of the models, this study will use overall accuracy, F1-score and FOM.

Planning and milestones

To make sure that on the third of December a complete and well written thesis is done the following milestones are planned. At the end of September, the data must be complete, preprocessed and ready to be trained upon. At the end of October, all programming must be finished and the models trained. Throughout these months, time will be spend on writing the thesis itself. At the half of November, the thesis must be finished and in principle be ready to turn in. Then the last two weeks are planned to be for finetuning the thesis.

References

Cadenasso, M. L., Pickett, T. A., & Schwarz, K. (2007). Spatial heterogeneity in urban ecosystems: reconceptualizing land cover and a framework for classification. *Front Ecological Environment*, *5*(2), 80–88. https://doi.org/https://doi.org/10.1890/1540-9295(2007)5[80:SHIUER]2.0.CO;2

Dembski, S., Hartmann, T., Hengstermann, A., & Dunning, R. (2020). Introduction enhancing understanding of strategies of land policy for urban densification. In *Town Planning Review* (Vol. 91, Issue 3, pp. 209–216). Liverpool University Press. https://doi.org/10.3828/tpr.2020.12

Du, G., Shin, K. J., Yuan, L., & Managi, S. (2018). A comparative approach to modelling multiple urban land use changes using tree-based methods and cellular automata: the case of Greater Tokyo Area. *International Journal of Geographical Information Science*, *32*(4), 757–782. https://doi.org/10.1080/13658816.2017.1410550

Kenniscentrum InfoMil. (n.d.). *Ladder voor duurzame verstedelijking*. Retrieved September 7, 2021, from https://www.infomil.nl/onderwerpen/ruimte/gebiedsontwikkeling/ladder-duurzame/

Ollongren, K. H. (2021, July 5). *Aanbieding Rapport Staat van de Woningmarkt 2021*. https://www.rijksoverheid.nl/documenten/kamerstukken/2021/07/05/aanbieding-rapport-staat-van-de-woningmarkt-2021

Omrani, H., Parmentier, B., Helbich, M., & Pijanowski, B. (2019). The land transformation model-cluster framework: Applying k-means and the Spark computing environment for large scale land change analytics. *Environmental Modelling and Software*, *111*, 182–191. https://doi.org/10.1016/j.envsoft.2018.10.004

Omrani, H., Tayyebi, A., & Pijanowski, B. (2017). Integrating the multi-label land-use concept and cellular automata with the artificial neural network-based Land Transformation Model: an integrated ML-CA-LTM modeling framework. *GIScience and Remote Sensing*, *54*(3), 283–304. https://doi.org/10.1080/15481603.2016.1265706

Shafizadeh-Moghadam, H. (2019). Improving spatial accuracy of urban growth simulation models using ensemble forecasting approaches. *Computers, Environment and Urban Systems*, *76*, 91–100. https://doi.org/10.1016/j.compenvurbsys.2019.04.005

Shafizadeh-Moghadam, H., Asghari, A., Tayyebi, A., & Taleai, M. (2017). Coupling machine learning, tree-based and statistical models with cellular automata to simulate urban growth. *Computers, Environment and Urban Systems*, *64*, 297–308. https://doi.org/10.1016/J.COMPENVURBSYS.2017.04.002

Wang, L., Omrani, H., Zhao, Z., Francomano, D., Li, K., & Pijanowski, B. (2019). Analysis on urban densification dynamics and future modes in southeastern Wisconsin, USA. *PLOS ONE*, *14*(3), e0211964. https://doi.org/10.1371/JOURNAL.PONE.0211964

White, R., Engelen, G., & Uljee, I. (1997). The use of constrained cellular automata for high-resolution modelling of urban land-use dynamics. *Environment and Planning B: Planning and Design*, *24*, 323–343. https://doi.org/doi:10.1068/b240323

Wolfram, S. (1983). Statistical mechanics of cellular automata. *Reviews of Modern Physics*, *55*(3). https://doi.org/10.1103/RevModPhys.55.601

Xing, W., Qian, Y., Guan, X., Yang, T., & Wu, H. (2020). A novel cellular automata model integrated with deep learning for dynamic spatio-temporal land use change simulation. *Computers and Geosciences*, *137*. https://doi.org/10.1016/j.cageo.2020.104430

Zhai, Y., Yao, Y., Guan, Q., Liang, X., Li, X., Pan, Y., Yue, H., Yuan, Z., & Zhou, J. (2020). Simulating urban land use change by integrating a convolutional neural network with vector-based cellular automata. *International Journal of Geographical Information Science*, *34*(7), 1475–1499. https://doi.org/10.1080/13658816.2020.1711915