Utrecht University

Applied Data Science Master's degree programme

Spatial Data Analysis and Simulation Modelling course

Topics and case studies

Document version: 1.2

Document modified: 2020.12.14

Dr. Simon Scheider

Department of Human Geography and Spatial Planning

Faculty of Geosciences

Contents

Topics

Active transport in a city

Introduction

Supporting active transport is of increasing importance for both fostering healthy urban living and physical activity of residents and for protecting the urban environment by reducing transport emissions. In many empirical studies, the effects of the environment on health or active behavior are investigated based on indices that measure opportunities for walking or biking. To come up with such an index, it is required to summarize the quality of the urban environment with respect to its walking or biking affordance, e.g., with respect to the accessibility of facilities and the quality of urban infrastructure.

Suggestions for short paper

You can research this topic in a short paper by discussing the quality and the approach of an existing study that makes use of such indices. You might also compare different studies/indices with each other. You should focus on three examples from literature, see suggestions below.

Suggestions for case study

In the case study, you should implement and compute some index to measure the walkability or the bikeability of some city (e.g. Amsterdam), based on existing geodata sources and based on existing models suggested from literature. For example, the suitability of a city for active transport (walking/biking) may be based on many indicators:

- Street network: Interconnectivity and quality of network for walking or biking
- Residential buildings: Origins of walking, density of buildings
- Greenness: walking experience, motivators for walking or biking
- Landuse: Motivators for walking, walking or biking experience
- Points of Interest: Destinations and distances for walking or biking

The goal of the study should be to compare active transport for different areas of the city or between cities. You should base your index on at least 3 of such factors, one for each person in your group, and thus based on 3 different data sources (see last section below) which need to be aggregated and transformed into the index. The index could be computed for different neighborhoods within the city or across different cities. Please add clear research questions, method arguments, references to model indices, map illustrations and comparable studies into your report.

Suggested data sources

You can use e.g. any of the datasets provided by the Amsterdam municipality:
https://maps.amsterdam.nl/ and https://data.amsterdam.nl/
In addition see PDOK:
https://www.pdok.nl/introductie/-/article/cbs-bestand-bodemgebruik
https://www.pdok.nl/introductie/-/article/basisregistratie-adressen-en-gebouwen-ba-1

Landelijk Grondgebruik Nederland (LGN)
https://www.wur.nl/nl/Onderzoek-Resultaten/Onderzoeksinstituten/Environmental-Research/Faciliteiten-tools/Kaarten-en-GIS-bestanden/Landelijk-Grondgebruik-Nederland.htm

Suggested literature

Frank, L. D., Schmid, T. L., Sallis, J. F., Chapman, J., & Saelens, B. E. (2005). Linking objectively measured physical activity with objectively measured urban form: findings from SMARTRAQ. American journal of preventive medicine, 28(2), 117-125.

Frank, L. D., Sallis, J. F., Saelens, B. E., Leary, L., Cain, K., Conway, T. L., & Hess, P. M. (2010). The development of a walkability index: application to the Neighborhood Quality of Life Study. British journal of sports medicine, 44(13), 924-933.

Van Dyck, D., Cardon, G., Deforche, B., Sallis, J. F., Owen, N., & De Bourdeaudhuij, I. (2010). Neighborhood SES and walkability are related to physical activity behavior in Belgian adults. *Preventive medicine*, *50*, S74-S79.

Winters, M., Brauer, M., Setton, E. M., & Teschke, K. (2013). Mapping bikeability: a spatial tool to support sustainable travel. *Environment and Planning B: Planning and Design*, *40*(5), 865-883.

Maghelal, P. K., & Capp, C. J. (2011). Walkability: A Review of Existing Pedestrian Indices. Journal of the Urban & Regional Information Systems Association, 23(2).

Koohsari, M. J., Owen, N., Cerin, E., Giles-Corti, B., & Sugiyama, T. (2016). Walkability and walking for transport: characterizing the built environment using space syntax. International Journal of Behavioral Nutrition and Physical Activity, 13(1), 1-9.

Yun, H. Y. (2019). Environmental factors associated with older adult's walking behaviors: A systematic review of quantitative studies. Sustainability, 11(12), 3253.

The potential for renewable energy

Introduction

One of the aims of the Dutch national government is to stimulate renewable energy in order to create a green and diversified energy system. For example, the province of North-Holland is obliged by the Dutch government to install 105.5 MW extra of onshore wind energy before 2020 (outside national wind park Wieringermeer). The province instructs you to assess the area suitable for onshore wind energy in the province of North-Holland. The aim is to understand how much space is available for this purpose.

Suggestions for short paper

You can research the assessment of renewable energy potential (wind energy or solar energy) in a short paper by discussing the quality and the approaches of existing studies. You can also compare different studies with each other in terms of their spatial analytic approach taken. You should focus on three examples from literature, see some suggestions below.

Suggestions for case study

Where in the province of North-Holland are locations suitable for building wind turbines? Suitability for windmill construction depends on a number of criteria and constraints. Since this is a preliminary study, at least 3 criteria and 3 constraints should be taken into account, which may be selected from this list:
Constraints:

- Not within 50 meters from roads, waterways and railtracks, and industrial pipelines
- not within nature areas (e.g. Natura2000)
- Not within 50 meters of built-up areas
- Not within 50 meters from recreational areas
- 200 meters from high voltage transmission lines or pipelines

Criteria:

- Large average wind speed
- Large distance from residential areas
- Large distance from nature areas
- Close to 50kV/150kV transformer station

The province wants to make a quick scan to understand if there is enough space for any turbines under these conditions, before ordering a full study and raising public fears or opinions. Your task is to assess this space in terms of a suitability map. Please add clear research questions, method arguments, map illustrations and comparable studies into your report.

Suggested data sources

Most data can be found on the website of the Province of North Holland: https://maps.noord-holland.nl/kaartenportaal/apps/MapSeries/index.html?appid=d8c0eb7751444d4b9537231615ad6a09

- Waterways: *Structuurvisie Basisnet Beroepsvaart/Basisrecreatietoevaartnet*
- Roads: *Structuurvisie regionaal wegennet*
- Railway: *Structuurvisie spoorweg*
- Nature: *Algemeen/Natura2000 gebieden, PRV Unesco gebieden*
- Built-up area: *Provinciaal Ruimtelijke Verordening/Bestaand Bebouwd Gebied (BBG)*
- High voltage net: http://www.hoogspanningsnet.com/netkaart/algemeen/
- Windspeed: https://globalwindatlas.info/area/Netherlands

Suggested literature

Denholm, P., Hand, M., Jackson, M., Ong, S. (2009). Land-Use Requirements of Modern Wind Power Plants in the United States. Technical Report NREL/TP-6A2-45834. National Renewable Energy Laboratory, Golden, Colorado, USA.

Grassi, S., Junghans, S., & Raubal, M. (2014). Assessment of the wake effect on the energy production of onshore wind farms using GIS. Applied Energy, 136, 827-837

Hoogwijk, M., Vries, B., de, Turkenburg, W. (2004). Assessment of the global and regional geographical, technical and economic potential of onshore wind energy. Elsevier Energy Economics. No. 26, pp. 889-919.

Mroczek, B., Kurpas, D., Karakiewicz, B. (2012). Influence of distances between places of residence and wind farms on the quality of life in nearby areas. Annals of Agricultural and Environmental Medicine 2012, Vol 19, No 4, 692-696.

Staatscourant (2002). Beleidsregel voor het plaatsen van windturbines op, in of over rijkswaterstaatswerken. Staatscourant, No. 123, pp. 13. https://zoek.officielebekendmakingen.nl/stcrt-2002-123-p13-SC35242.html

Rijksoverheid (2020): Windenergie op land https://www.rijksoverheid.nl/onderwerpen/duurzame-energie/windenergie-op-land

Ramachandra, T. V. (2006). Solar energy potential assessment using GIS. *Energy Education Science and Technology*, *18*(1/2), 101.

Groppi, D., de Santoli, L., Cumo, F., & Garcia, D. A. (2018). A GIS-based model to assess buildings energy consumption and usable solar energy potential in urban areas. *Sustainable cities and society*, *40*, 546-558.

Automating meta-data descriptions for geodata sources

Introduction

The availability of meta-data for geodata sources is essential to assess whether a given data source is useful for data science. Manual meta-data descriptions, however, are cumbersome to generate and do not scale with the increasing amount of data available on the Web. The problem of automating meta-data descriptions therefore becomes more and more important. Automating the description of geodata sources with meta-data, however, can be very challenging or even impossible, depending on what kind of information is inferred. Nevertheless, some meta-data can be automatically extracted from the data. We are mostly interested in core concept data types, measurement levels, but also geodata types and geodata quality measures (resolution, extent, completeness, accuracy) would be useful. How could we solve this problem programmatically, at least for some aspect of meta-data, and to what extent can meta-data annotation be automatized?

Suggestions for short paper

You should review the state-of-the-art of studies on meta-data inference for geodata sources. Compare different approaches and discuss the limitations and main challenges of existing approaches (see references below).

Suggestions for case study

You should develop and test a script that tries to infer some meta-data labels based on a given set of geodata sources provided below. Labels could be core concept data types, measurement levels, but also geodata types or geodata quality measures (resolution, extent, completeness, accuracy) and the labeled dataset is given as a sample (see below). Analysis should be done based on extracting certain features from the geodata sources, such as geometry types or text descriptions, and using some Machine Learning (classification) or some rule based model to test whether you can find ways to label data automatically. Train your classifier and measure its quality based on manually added labels. You can also do a more exploratory study, in which you try to interpret rule models like decision trees, frequent itemsets or clusters according to whether they are able to distinguish the class labels in a meaningful way. The goal is to find out which meta-data aspect might be suitable for automation, and which ones are rather difficult to automize. Please add clear research questions, method arguments, illustrations and comparable studies into your report.

Suggested data sources

We provide a dataset of geodata sources annotated with core concept data (CCD) types for the *Amsterdam Data Portal (*https://maps.amsterdam.nl/open_geodata/*)*. An annotated snapshot of the datasets from the Portal can be downloaded from the Google Folder at https://drive.google.com/drive/folders/13RncYOD9_MV77VML-vR4Bs1LJXE5UFKz?usp=sharing . Each dataset in the snapshot was annotated in terms of core concepts and attribute

measurement scales, using classes from the CCD ontology (Scheider et al. 2020) (http://geographicknowledge.de/vocab/CoreConceptData).

Suggested literature

Bernard, L., *et al*., 2014. Scientific geodata infrastructures: challenges, approaches and directions. *International Journal of Digital Earth*, 7 (7), 613–633

Egenhofer, M.J., 2002. Toward the semantic geospatial web. *In*: *Proceedings of the 10th ACM international symposium on Advances in geographic information systems*

Manso-Callejo, M., Wachowicz, M., & Bernabé-Poveda, M. (2010). The design of an automated workflow for metadata generation. In *Research Conference on Metadata and Semantic Research* (pp. 275-287). Springer, Berlin, Heidelberg.

Klien, E. (2007). A Rule-Based Strategy for the Semantic Annotation of Geodata. *Transactions in GIS*, *11*(3), 437-452.

Trilles, S., Díaz, L., Gil, J., & Huerta, J. (2012). Assisted generation and publication of geospatial data and metadata. *International Journal of Spatial Data Infrastructures*, 24-27.

Scheider, S., Meerlo, R., Kasalica, V., & Lamprecht, A. L. (2020). Ontology of core concept data types for answering geo-analytical questions. *Journal of Spatial Information Science*, *2020*(20), 167-201. (http://www.josis.org/index.php/josis/article/viewArticle/555)

Scheider, S., & Huisjes, M. D. (2019). Distinguishing extensive and intensive properties for meaningful geocomputation and mapping. *International Journal of Geographical Information Science*, *33*(1), 28-54.

Land use change modelling using cellular automata

Introduction

Cellular automata use local neighbourhood interactions to simulate larger scale behaviour of a spatio-temporal system. These local neighbourhood interactions are given by transition rules valid for each cell, where the state of a cell changes over a timestep as a function of the state of cells of directly neighboring cells. Cellular automata are widely used, e.g. for modelling plant growth and spread, modelling forest fire spreading, modelling growth of bacteria on leaves of vegetation, modelling socio-economic systems. The aim of this topic is to learn more about cellular automata modelling by studying the use of this type of models for forecasting land use change.

Suggestions for topic paper assignment

The topic is on the use of cellular automata to model land use change. You could describe the concepts of cellular automata for modelling land use change, possibly describing other approaches as well, for instance agent-based type of techniques, where individual decisions of farmers are included. If you wish you can also focus on methods to tightly integrate observational data on land use and land use change models You can then use one or two examples from the literature to illustrate the concepts.

Suggestions for topic case study

You will construct a simple cellular automata model of expansion of cities (Randstad, the Netherlands). Start with the land use situation in 2000 (as represented by the data set) and try to simulate the change in landuse over the coming decennia. Use simplified land use change transition rules that can be implemented in the framework of this short case study. Be sure to give a clearly defined objective of the study and 1-2 research questions. The aim cannot be solely to construct the model, the model has to be a means to answer a particular question.

Download dataset and model. Download information regarding the data set.

Suggested literature

Torrens, P.M., 2000, How cellular models of urban systems work (1. theory). Centre for advanced spatial analysis, working paper series. Paper 28. Available at http://www.bartlett.ucl.ac.uk/casa/publications/working-paper-28

Batty M., Xie Y., Sun Z., 1999. Modeling urban dynamics through GIS-based cellular automata. Computers, Environment and Urban Systems 23:205-233. To retrieve this paper, email me and I will email you the pdf.

Integration of data and simulation models for forecasting

Introduction

Simulation models are widely used for forecasting of spatio-temporal systems. Examples are hydrological models, epidemiological models, climate models. To minimize forecast uncertainty it is essential to integrate observational data with simulation models. One approach is model calibration, which involves the adjustment of model parameters to better match model outputs with observational data. In this topic you will study the use of calibration in simulation modelling and execute a case study on model calibration.

Suggestions for topic paper assignment

In the first part of the paper describe a particular concept from model calibration. An obvious choice would be to describe one or a group of calibration algorithms, for instance hillclimbing, genetic algorithms, Bayesian methods. In the second part of the paper, then, illustrate these concepts by describing model calibration in one or two case studies. Case studies considered can be from any domain.

Suggestions for case study

The case study will be done using the same model and model area as considered in the computer labs. There are multiple paths that you can follow in the case study:

1. Improve upon the calibration by calibration of multiple (all?) parameters in the model. Use the brute force technique and code it yourself (using Python or numpy) by extending the runoff.py script.
2. Same as above but using SPOTPY (https://pypi.org/project/spotpy/), which provides built-in calibration routines. This will require you to write code that calls the model from SPOTPY Python functions.
3. Extend your analysis towards an independent validation period. That is, calibrate the model for a certain time span (e.g. year 1-3) and validate (test) it for another time span ('prediction'), e.g. year 4-6. For this you need to extend the time period you run the model, this can be done with the second download below. You may want to stick to calibrating one or maybe two parameters.
4. Compare the capability of the model to predict discharge with a statistical learning algorithm. An alternative to using the simulation model (and calibration) is to train a statistical learning model (e.g. regression, random forest, deep learning) on the data (it should be able to predict streamflow from precipitation and temperature alone without using any other model or the map data). Test this approach and compare the results with those from calibration of the simulation model. Note that it is preferable to separate between training (calibration) and validation (testing) time periods, like 3) above.
5. The current data set used for meteorology is somewhat old. A more recent reanalysis data set with meteo data is ERA5 (https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5). The idea here is to compare model runs that rely on this alternative meteo input. You need to

download the ERA5 data for the correct location of the catchment and resample it to a 1 day time step (the download is in 1 h timesteps). And then run the model with the same data (and possibly calibrating it again).

Data set

Use the same data set as used in the labs, you can download it from
http://karssenberg.geo.uu.nl/lspm/_static/calibration.zip

If you wish to extend or change the time period used for calibration (or validation), download the source data for observed streamflow and meteorology. The download includes a Python script to extract the data that you wish to use and converts to a format that can be read by the simulation model. Use this download:
http://karssenberg.geo.uu.nl/lspm/_static/observed_meteo_and_discharge.zip

Be sure to read the readme.txt document with information on the content of the downloads.

Suggested literature

Start from the literature that you studied for the exam (see Blackboard) and papers that refer to this literature. A search on calibration methods in for instance scopus will also give you sufficient literature.

Agent-based modelling in epidemiology

Introduction

Agent-based models are widely used in epidemiology. Unlike purely statistical or machine learning approaches they provide understanding of mechanisms that lead to dispersal of diseases or certain related phenomena like dietary habits, as you studied in the agent-based modelling computer labs during our course. In this topic you will study the use of agent-based modelling epidemiology.

Suggestions for topic paper assignment

There is of course a wealth of literature on agent-based modelling in epidemiology. The paper should describe a particular concept and then illustrate this concept with one or two case studies from epidemiology. Suggestions for the concept are: 1) to focus on a particular disease (e.g. vector borne disease, cardio-vascular disease, COVID-19) and describe agent-based modelling concepts in this domain, 2) to focus on a particular challenge in agent-based modelling of disease in general, e.g. usage of data in agent-based models, either as model input, or for calibration of models, 3) to focus on uncertainty in the outcomes of agent-based models, and how errors in inputs or model concepts propagate to the model output.

Suggestions for topic case study

The case study will build upon the agent-based model that you constructed during the agent-based modelling labs. You will execute multiple scenarios that refer to particular assumptions on dietary behaviours or on how food outlets respond to food demands from their customers. You can follow one (or a combination of) the paths below:

1. Study the effect on changes of propensity for health food (and resulting spatial patterns) of the distance travelled by persons to buy food, i.e. change the buffer size.
2. Study the effect of changes in other model parameters on the model outcome. Define scenarios of model parameter values and run the model for these scenarios.

Data set

Same as used for the labs.

Suggested literature for the case study (mainly dietary habits related)

Orr et al., 2017, http://dx.doi.org/10.1136/jech-2015-205621

Auchincloss et al., 2008, https://doi.org/10.1093/aje/kwn118

Potential data sources

- Maps Amsterdam:
  https://maps.amsterdam.nl/
- Data Amsterdam:
  https://data.amsterdam.nl/
- Nationaal Georegister:
  http://www.nationaalgeoregister.nl/geonetwork/srv/dut/catalog.search#/home
- PDOK:
  https://www.pdok.nl/
- Portaal Kaart en Data Noord Holland:
  https://maps.noord-holland.nl/kaartenportaal/apps/MapSeries/index.html?appid=d8c0eb7751444d4b9537231615ad6a09
- Mapcruzin:
  https://mapcruzin.com/download-free-arcgis-shapefiles.htm
- Statistics Netherlands (CBS):
  https://opendata.cbs.nl/statline/#/CBS/nl/
- OpenStreetMap::
  https://www.openstreetmap.org/
  http://overpass-turbo.eu/s/LpE
  http://openpoimap.org/
- Joint Research Center Data Catalogue:
  https://data.jrc.ec.europa.eu/
- Route Databank:
  https://www.routedatabank.nl/
- Copernicus Open Access Hub:
  https://scihub.copernicus.eu/