OPEN DATA AND TRANSPORT MODELLING

Dmitrii Grishchuk*,[1]

*222298@vutbr.cz
[1]Brno University of Technology, Faculty of Civil Engineering, Veveří 331/95, 602 00 Brno, Czech Republic

Abstract

Computer modelling has become one of the main methods of predicting the behaviour of traffic and passenger flows within small areas, cities, or entire regions. Sometimes the lack of data, either open or proprietary, or the lack of knowledge of where to find it and how to prepare it, is an obstacle to the development of a traffic model.

In this paper, the author has chosen to focus on finding and processing the freely available data on the population, infrastructure, facilities characteristics and placement for developing traffic models in any locality in the Czech Republic, regardless of size. The format of the processed data is chosen in such a way that it can be used for other analytical purposes.

Keywords

Open-source, data, transport modelling, mobility, points of interest

1 INTRODUCTION

Predictions based on mathematical models are very helpful means of foreseeing the possible outcome of certain events. However, sophisticated models tend to require large amounts of precise data that are often hard to obtain. Models used in transport simulations belong to this group as well. Hence, a researcher willing to obtain reasonable results out of these models should make sure that inputs meet with the required level of quality.

The open data topic has grown relevant in the last decade, and many world governments, among others the European Union, have issued directives encouraging or even obliging the public sector to make their collected data widely and freely available [1], [2]. The objectives of these initiatives incorporate the provision of high-quality data for scientific and other activities for public benefit, including transport models' creation.

There are several approaches to handle inputs in the field of traffic modelling. Sometimes algorithms rely on the use of statistical zones of various sizes with different production and attraction characteristics [3]. The output of such an algorithm is typically presented as a count of trips between each pair of zones per day which gets utilized during traffic or passenger load assignment.

The other way is to work with precise locations during the demand creation, which is especially useful when combined with the concept of agents, i. e. synthetic representations of individuals with certain attributes and distinct daily trips chains, like in MATSim [4]. Depending on the actual implementation, this allows for watching every agent's position at any given moment.

Regardless of the method, the data for the models must be secured, and the more precise they are, the wider is the set of use cases.

The data for transport models usually consist of:

- networks with parameters, e.g. road or rail,
- zones or stand-alone facilities with parameters for trips creation,
- statistical overviews for population in zones,
- software-specific configuration files.

In this respect, the paper focuses on sources of open data and on ways of their acquisition and conversion into formats suitable for transport modelling. The primary region of interest is the Czech Republic; however not all the sources are exclusive to it and can therefore be used in other regions.

2 SOURCES AND METHODOLOGY

One of the best-known sources of open spatial data is OpenStreetMap (OSM) [5]. It is built by its community of contributors and features elements usually contained by mapping services: networks, buildings geometries, points of interest, land use and more. There are several ways to retrieve data in the OSM formats. These include services

DOI 10.13164/juniorstav.2024.24009

that allow downloading the current OSM data, namely the HOT Export Tool [6] and the BBBike [7]. Some organizations save the OSM snapshots – a well-known tool for this purpose is Planet OSM [8]. The Faculty of Information Technology of the Brno University of Technology makes a daily back up of the OSM data in the PBF (protocol buffer) format for the whole of the Czech Republic on its servers [9].

There are also other existing sources for specific types of data that will be discussed in the corresponding sections below.

Network

All transport simulation software requires a road network to perform traffic and passenger load assignment. Multi-modal suites additionally accept alternative networks, e.g. for trams, trains, metro and other means of transport with partially or fully isolated infrastructure.

The official network data for the Czech Republic are provided by the State Administration of Land Surveying and Cadastre.

One of the datasets is called Data50 and contains the ESRI shape files with line geometry describing the individual parts of networks – railways, highways, streets, bridges, tunnels, and some of the pedestrian infrastructure [10]. However, it lacks information on speed restrictions, line directions, the width or the number of lanes (tracks), which are crucial parameters for traffic dynamics simulation. Moreover, there is no information about tram tracks that may be necessary for certain scenarios.

Another dataset is ZABAGED (Czech abbreviation for The Fundamental Base of Geographic Data of the Czech Republic) distributed in several parts as geopackage files [11]. It offers more detailed network information than the previously described dataset except for the speed limits on links.

To the author’s best knowledge, there are no publicly available tools that allow for the conversion of the aforementioned data types into files accepted by any transport modelling software. What is known to be supported to some extent, is the OSM data format. Its network contains even more information, including that on speed on network links and turn restrictions on intersections. Nevertheless, compared to the official sources, the OSM has a drawback – because of being built by a community, the data quality may not be consistent in various areas which complicates the conversion.

Proprietary software products, such as PTV Visum and Aimsun, usually have built-in import functions [12], [13]. For open-source software, there are plug-ins or scripts helping to convert original data into native formats, most notably netconvert and osmWebWizard for SUMO [14] and SupersonicOsmNetworkReader for MATSim [15]. However, the MATSim’s tool is not ready-to-use as it requires configuring the MATSim environment first.

In many of the tools, the hour capacity is not considered during conversion or it is assigned some constant uniformly, e.g. 1800 vehicles per hour per lane, regardless of road type and speed limit. The author of this paper proposes a ratio (1) for estimating capacity:

$$C_e = \min(17{,}5 \cdot \upsilon_e \cdot \mathrm{k}_e + 60{,}1800) \cdot N_l \qquad (1)$$

where C_e is the edge capacity, υ_e is the edge speed limit, k_e is the edge type coefficient and N_l is the edge lane count. The edge type coefficient k_e varies for different types of roads, as shown in Tab. 1.

Tab. 1 Default values for selected OSM link types.

Edge type	Edge type coefficient k_e	Edge lane count N_l	Edge speed limit υ_e, km/h
motorway / motorway_link	*1.1*	*2*	*130 / 80*
trunk / trunk_link	*1*	*2*	*110 / 80*
primary / primary_link	*1*	*2*	*90 / 70*
secondary / secondary_link	*0.95*	*1*	*60 / 50*
tertiary / tertiary_link	*0.9*	*1*	*50 / 50*
residential	*0.8*	*1*	*30*
living_street	*0.7*	*1*	*20*

The ratio (1) is designed primarily for MATSim because of the way it handles capacity and trips assignment. Less significant road links get lower hour capacity, which discourages the agents from using it once they “realize” that the capacity is not sufficient for just passing by and they will traverse it only when necessary.

DOI 10.13164/juniorstav.2024.24009

Points of interest

Points of interest (POIs) are also referred to as facilities in some terminologies. They represent places where different types of activities are performed. For transport models based on statistical zones points of interest are aggregated.

The most comprehensive source for this kind of data is OpenStreetMap. However, depending on the point type, the information provided might not be sufficient.

For instance, the data about residential buildings do not contain the number of inhabitants. The only city in the Czech Republic that was found to provide the number of residents per address was Brno [16]; thus, another source must be utilized when dealing with other areas. The author of this paper proposes a universal method of assigning residents to buildings. First, an estimate of the number of residents is made, based on the total gross floor area (gross single floor area multiplied by the number of floors) according to the ratios derived from a Czech certified methodology for estimation of traffic potential [17]: 20 m^2/person for dormitories, 50 m^2/person for general residential and apartment buildings, and 80 m^2/person for detached houses. Then the estimate is corrected with regard to the population per basic settlement units (BSU) from the 2021 Census data [18]. The estimate of the number of residents in buildings within a BSU is compared with the actual data, and in case of a difference, the number of residents is proportionally reduced or increased based on the gross total floor area of each building. This step is unnecessary in case of aggregation to the level of BSU, but is mandatory for agent-based models.

One of the most fundamental aspects affecting people's mobility is work. Locations of workplaces cannot be estimated as easily as the number of residents since the Census data do not provide information about the number of employees with any precision. For getting workplaces on addresses, the author of this paper proposes to use two sources that can, when combined, provide a relatively precise estimate of workplaces on addresses. The Register of Economic Entities [19] lists all companies registered in the Czech Republic and contains an approximate employees count of the whole organization, while the Trade Licensing Register provides an interface [20] to query a database for getting all the organization's branches. The total number of an organization's employees can be then divided by the number of its branches. As there are no spatial data in any of the used sources, addresses must be geocoded in order to locate their coordinates.

Trips for education contribute noticeably to younger generation's mobility. The existing Register of Educational Institutions [21] provides addresses and capacities of all education-related organizations; however, it does not provide an option to bulk download data. Manual selecting and copying is out of the question due to significant effort and time expenses; an automated tool would be a more rational way of acquiring the data.

Shopping is a massive part of mobility and should also be addressed in transport models. Visitor counts depend on the assortment of goods and the shop's area. Shops may be partially extracted from OSM, though because of its specificity in points of interest representation, many businesses are kept as nodes rather than polygons, thus, their area cannot be estimated. In such cases manual data revision is necessary. The city of Brno is an exception because it provides data coming from retail research [22] which takes place in the city once in several years.

POIs for other activities (leisure, errands) are optional, but can improve outputs quality, when there's enough statistical data.

Statistics

For definition of behavioural characteristics of a population, censuses and surveys are conducted. The authorities then publish the results, which may contain processed data with a certain level of detail for further analyses that the general public can extract insights from.

The most recent census in the Czech Republic was carried out in the year 2021 [18]. It included information about population size and distribution, people's education, occupation, housing and more, with different levels of spatial detail. Even though it is possible to extract basic daily mobility patterns out of this source, the available data might not cover all the needs for a transport model – the 2021 Census only provides work and school commuting data. For transport models it is also important to know how far people travel and at what time. An answer to this problem is a specialized mobility survey. Such a study was conducted from 2017 to 2019 by the Transport Research Centre and is called The Czech Republic in Motion [23]. As opposed to the Census data, its results contain anonymized daily travel diaries of real people which include the purpose of the trips (activities), lengths, modes, start and end times. Spatial precision for origins and destinations is on the level of municipalities with extended competence. The interviewed people are described from the social and demographic point of view, i. e. in terms of their sex, age, education, occupation, car ownership etc. This enables for mobility index calculation or reusing travel diaries to specify the synthetic population's behaviour in a transport model. Population social and demographic categories ratios for the used zones are available in the Census data.

3 RESULTS

To help transport model creators, possible open data sources were aggregated and discussed in this work.

The author of this paper has also created a tool in the Python programming language that extracts network from the OSM protocol buffer into MATSim network XML file [24], the ratio (1) and default values from Tab. 1 were used. The tool fills in missing speeds based on link's category and estimates hour vehicle capacities. It can also output the ESRI shape files for a user to be able to check the exported OSM attributes graphically using one of the GIS programs. The tool does not take turn restrictions into account yet, but this functionality will be implemented in the future. An example of a graphical representation of such a network is shown in Fig. 1.

Another tool prepared by the author is dedicated to downloading and aggregating data about Czech educational institutions from the Register of Educational Institutions [25]. Additionally, the data get geocoded by each institution's address to include its location. The already processed and manually corrected data are bundled with the tool and can be downloaded separately. The data, however, are not ready for use in a transport model as they are and need to be converted according to the target software's specifications. A graphical representation can be seen in Fig. 2.

Tools for other points of interest (homes, workplaces, shopping centres/shops, leisure facilities, errands) are currently being developed but cannot be published yet because of their incompleteness.

4 DISCUSSION

The completed research has shown that there are enough sources of open data for the creation of transport models for regions or cities of the Czech Republic. Although data from some sources cannot be obtained conveniently, the issue is solvable with the usage of programming tools. Some of them were and are being developed by the author of this paper but the time and resource complexity is not optimized for the use on average tier computer hardware: scripts can take up to 16 gigabytes of random-access memory (RAM) and several hours of run time. The tools were developed using the Ubuntu 22.04.2 operating system (Linux) and were not tested on other more popular operating systems – Windows and macOS. Moreover, it is known that the tool for educational institutions download will not work on Windows since it uses the Python's multiprocessing library differently. It requires that when run on Windows, the function being passed to a multiprocessing pool is imported from another file or module, while the actual tool imports from within the current file.

5 CONCLUSION

The results of the conducted open data sources research have proven that the amount of open data for transport models is satisfactory in the Czech Republic. Not all data sources had a convenient way of obtaining their contents, and the process had to be done using programming tools which are not all done yet. Further directions of the work are:

- regular search for new suitable open data sources,
- continuation of the development of other programming tools for open data collection and conversion,
- optimizing the already developed tools in terms of consumed system resources and time,
- adapting and testing of the developed tools on Windows and macOS.

DOI 10.13164/juniorstav.2024.24009

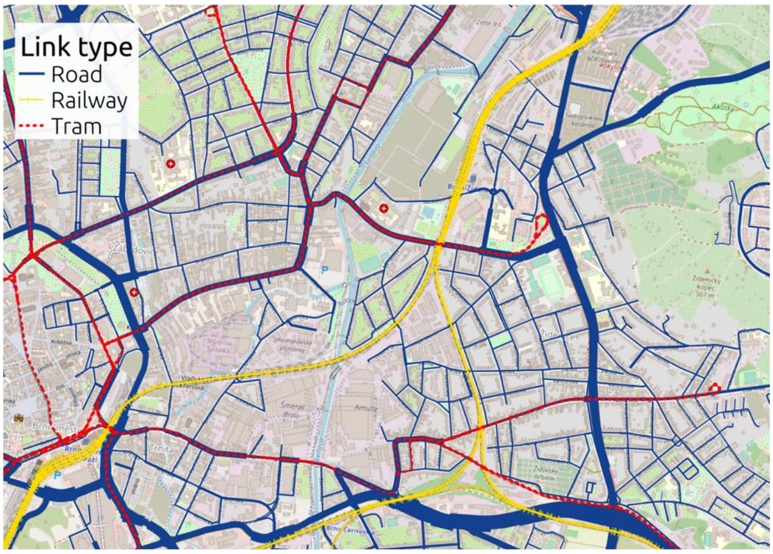

Fig. 1 Graphical representation of an exported OSM network in QGIS, widths of road links are relative to their capacities.

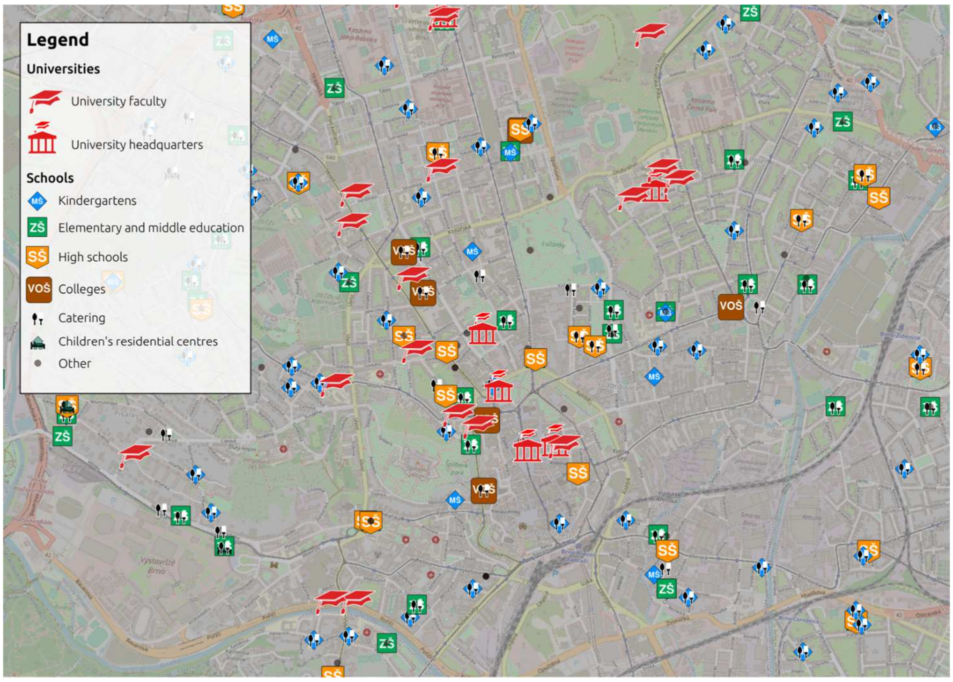

Fig. 2 Graphical representation of educational institutions in QGIS.

References

[1] NIKIFOROVA, Anastasija. *Smarter Open Government Data for Society 5.0: Are Your Open Data Smart Enough?* Sensors. 31 July 2021. Vol. 21, no. 15, p. 5204. DOI 10.3390/s21155204

[2] THE EUROPEAN PARLIAMENT. *PE/28/2019/REV/1 Directive* [online]. *Official Journal of the European Union*. 2019. [Accessed: 2023-11-11]. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32019L1024

[3] PATRIKSSON, Michael. *The Traffic Assignment Problem: Models and Methods*. Dover ed. United States: Courier Dover Publications, 2015. ISBN 978-0-486-78790-9

[4] HORNI, Andreas; NAGEL, Kai and AXHAUSEN, Kay. *Multi-Agent Transport Simulation MATSim*. London: Ubiquity Press, 2016. ISBN 978-1-909188-76-1. DOI 10.5334/baw

[5] OPENSTREETMAP FOUNDATION. OpenStreetMap [online]. United Kingdom: Openstreetmap Foundation, 2004. [Accessed 2022-12-03]. Available at: https://www.openstreetmap.org

[6] HOT AND FRIENDS. *HOT Export Tool* [online]. Available at: https://export.hotosm.org/en/v3/. [cit. 2023-11-13]

[7] SCHNEIDER, Wolfram. *BBBike exports of OpenStreetMap data (OSM, Garmin, mapsforge etc.)* [online]. Bbbike.org. [accessed 2023-11-13]. Available at: https://download.bbbike.org/osm/

[8] OPENSTREETMAP FOUNDATION. *Planet OSM* [online]. Openstreetmap.org. [accessed 2023-11-13]. Available at: https://planet.openstreetmap.org/

[9] FACULTY OF INFORMATION TECHNOLOGY BUT. *Index of /extracts/czech_republic* [online]. Osm.fit.vutbr.cz. Available at: https://osm.fit.vutbr.cz/extracts/czech_republic/. [cit. 2023-11-13].

[10] STATE ADMINISTRATION OF LAND SURVEYING AND CADASTRE. *Data50*. Online. Geoportal ČÚZK. 2023. [accessed 2023-11-12]. Available at: https://geoportal.cuzk.cz/(S(ssfxuhrr25v0muv4iqfvzn4s))/Default.aspx?lng=EN&mode=TextMeta&side=mapy_data50&text=dSady_mapyData50&head_tab=sekce-02-gp&menu=2290

[11] STATE ADMINISTRATION OF LAND SURVEYING AND CADASTRE. *ZABAGED® – planimetric components – introduction* [online]. Geoportal ČÚZK. 2023. [accessed 2023-11-12]. Dostupné z: https://geoportal.cuzk.cz/(S(gcpn3kil22ubd15qwfzyhkfy))/Default.aspx?lng=EN&mode=TextMeta&text=dSady_zabaged&side=zabaged&menu=24

[12] AIMSUN SLU. OpenStreetMap Importer [online]. Aimsun Documentation. 2022. [accessed 2023-11-12]. Available at: https://docs.aimsun.com/next/22.0.1/UsersManual/OSMImporter.html

[13] PTV MOBILITY. Webinar: From OSM via PTV Visum to PTV Vissim [online]. *YouTube*. 2017, [accessed 2023-11-12]. Available at: https://www.youtube.com/watch?v=m2Ol79VCDu4

[14] DLR INSTITUTE OF TRANSPORTATION SYSTEMS. OpenStreetMap [online]. *SUMO User Documentation*. 2023. [accessed 2023-11-12]. Available at: https://sumo.dlr.de/docs/Networks/Import/OpenStreetMap.html

[15] MATSIM ASSOCIATION. SupersonicOsmNetworkReader [online]. *GitHub*. 2023, [accessed 2023-11-12]. Available at: https://github.com/matsim-org/matsim-libs/tree/master/contribs/osm

[16] STATUTORY CITY OF BRNO. Number of people living at the addresses [online]. Data.brno.cz. 2022, [accessed 2023-11-13]. Available at: https://data.brno.cz/datasets/89d09657b1464911a195249d18610677_0/explore

[17] MARTOLOS, Jan et Al. Metody prognózy intenzit generované dopravy [online]. 1. Praha: *EDIP*, 2012, [accessed 2023-11-13]. Available at: https://www.mdcr.cz/getattachment/Dokumenty/Veda-a-vyzkum/Certifikovane-metodiky/Ostatni-metodiky/Metody-prognozy-intenzit-generovane-dopravy/Metody-prognozy-intenzit-generovane-dopravy.pdf.aspx

[18] CZECH STATISTICAL OFFICE. Výsledky sčítání 2021 – otevřená data [online]. *Czech Statistical Office*. 2022, [accessed 2023-11-13]. Available at: https://www.czso.cz/csu/czso/vysledky-scitani-2021-otevrena-data

[19] CZECH STATISTICAL OFFICE. Registr ekonomických subjektů: otevřená data [online]. Czso.cz. Prague: *Czech statistical office*, 2022, [accessed 2023-11-13].Available at: https://www.czso.cz/csu/czso/registr-ekonomickych-subjektu-otevrena-data

[20] MINISTRY OF INDUSTRY AND TRADE. Subject search [online]. *Rzp.cz*. 2017, [accessed 2023-11-13]. Available at: https://www.rzp.cz/cgi-bin/aps_cacheWEB.sh

[21] MINISTRY OF EDUCATION, YOUTH AND SPORTS. Rejstřík škol a školských zařízení [online]. *Msmt.cz*. 2004, [accessed 2023-11-11]. Available at: https://rejstriky.msmt.cz/rejskol/default.aspx

[22] STATUTORY CITY OF BRNO. Brno retail research [online]. *Data.brno.cz*. 2021, [accessed 2023-11-13]. Available at: https://arcg.is/1KDTu0

[23] KOUŘIL, P., ŠIMEČEK, M, DYTRT, Z. Česko v pohybu. Metoda a základní výsledky celostátního průzkumu dopravního chování. *Centrum dopravního výzkumu*, v. v. i., ISBN 978-80-88074-96-0

[24] GRISHCHUK, Dmitrii. Osm_net_to_matsim [online]. *GitHub*. 2023, [accessed 2023-11-25]. Available at: https://github.com/leonefamily/osm_net_to_matsim/tree/main

[25] GRISHCHUK, Dmitrii. Czech_educational_institutions [online]. *GitHub*. 2023, [accessed 2023-11-25]. Available at: https://github.com/leonefamily/czech_educational_institutions/tree/main