ORIGIN DESTINATION MOBILITY FLOW INFERENCE FROM TRAFFIC DATA

Case study on 2022 Hungarian public road data

By
Mihaly Hanics

Central European University
Department of Network and Data Science

Research internship report as requirement for the degree of Master of Social Data Science

Supervisor: Márton Karsai

Vienna, Austria
2024

CONTENTS

1. INTRODUCTION

1.1. General motivation

For the past century, (quantitative) human mobility analysis has been key not only for understanding mobility and its patterns, but also to analyze other social phenomena which are sensitive to society level mobility proportions. In some fields and areas, measuring some form of human mobility is essential for accurate outcomes, such as traffic and traffic congestion prediction, or long-term urban planning. Traffic prediction and immediate urban decision making usually require real-time data or at the very minimum frequent updates on current commuting volumes. Long-term planning in most cases only requires temporal averages over a long period. In other areas, the analysis or inferred measures are not directly or easily measurable, quantifiable, such as social interactions for predicting epidemic spreading, or social behaviour. In such cases, a useful predictor, or a highly correlated and easily measurable proxy would be the volumes or proportions of human mobility between predetermined, relevant locations. While this may not be sufficient to make individual predictions on local interaction level (unless tracking individuals), it is generally a helpful measure for higher level observations, such as an estimate of how many people will get infected in a location the upcoming day. In fact, such methods were developed also during the COVID-19 outbreak to assist decision-making and derive conclusions: among these, two common approaches (for estimating mobility sizes, not for epidemy analysis) are key for our work: the typically simpler and more accessible method of using population distributions to estimate mobility sizes[i] [ii], and directly measuring commuting between relevant locations[iii] [iv]. Some additional methods include using self-reported or collected contact data[v], and cellular phone data for real-time tracking[vi]. In our research, we would like to infer mobility volumes, mobility flow patterns between Hungarian settlements. We take inspiration from and compare to the population-size based inference methods, but primarily work with and focus on commuting-data-based methods. Our main result would be to create a table of mobility volumes between origin locations and destination locations, namely the origin-destination matrix. We create this matrix including all major traffic hubs of Hungary, and we further aim to include as many settlements as possible, while still including only proportionally correct values to ensure accuracy. For Hungary, constructing an O-D matrix from traffic data can improve the accuracy of previous methods while keeping a high-resolution, which in turn can provide better basis for epidemic spreading models, and provide unseen insights into commuting behaviours for better infrastructure and urban planning (optimize traffic flows, reduce travel times).

1.2. Problem description

In this particular work, we aim to reconstruct the O-D matrix for major Hungarian cities and popular destinations, leveraging traffic information to improve on current results, helping subsequent research and decision making, planning. An O-D matrix consists of pairs of locations, where each pair represents the number of people traveling from one location (the origin) to another (the destination) over a specified time period, such as how many people commute from Budapest to Szeged on an average day. The challenge in creating such a matrix lies in accurately capturing the number of people commuting between these locations on a regular basis, such as daily commuters. Traditional data collection methods, like surveys, are typically conducted infrequently and unless very extensive, cannot reliably be used to estimate accurate proportions of commuters between locations. Moreover, any rarely connected data cannot capture dynamic, “constantly changing” nature of temporal commuting patterns (for an example, many workers commute from nearby towns to Budapest in the morning, and travel home only in the evening, creating an asymmetric pattern). For real-time use, these can provide a baseline, but not for change detection. Mobile phone data can be used when real-time tracking is required, however is generally hardly accessible (network providers-maintainers often hesitate to provide researchers with sensitive data), raises privacy concerns, further requiring strong commitment from both parties to ensure mature procedures and protocols – even if this establishment between the two parties is successfully created, it may be only for short-term, results might be published only with restrictions. Technically, it also comes in a more complex type of data structure.

These limitations motivate traffic data to also be used for models, which is usually one of the finest approaches to infer mobility flows among the widely available data sources. It aids the issue of dynamic changes by being available at any times with frequent updates from the measurement devices on the roads, and also doesn’t raise any of the concerns stated before. The data is collected on public roads and no sensitive information is stored (vehicles are not tracked and certainly not through multiple roads, only vehicle counts are stored) and is also not provided by profitmaking private companies who do collect the data as part of their commercial service, but provided by governmental organizations who risk less by providing data for researchers, therefore typically are more open to do so.

In our case, we are in no need of dynamic mobility flow prediction, we only work with monthly and annual average commuting, but for all the other reasons and to improve on current methods we mainly work with traffic data for inferring mobility flows. From the measured traffic on the road, we use models with heuristics to infer how many people commute between a pair of settlements. This is done for all pairs of locations that we work with, which are initially major traffic hubs of Hungary, and in later work we aim to do the same for all settlements above 1000 population. We also evaluate and analyze our results, including comparing to a ground-truth-like dataset.

1.3. Challenges

We have already mentioned the issues with other methods for origin-destination flow inferring that we do not work with. We have also mentioned that solely population-based models are good baselines that we want to improve on (data for these models is usually available for wide ranges). There are some issues and shortcomings of working with models that use traffic data for inference or prediction; some are more general problems, some are specific for our case. The main general problem is that even with "perfect" data, one cannot theoretically reconstruct the O-D matrix in vastly all cases simply from traffic data – this is explained more in the respective chapter. Technically, accuracy of the reported data may be also an issue, and errors can vary highly depending on the road traffic measurement systems deployed (in our case, for each road segment, the error ranges are reported). A difficulty in our case is the geometric nature and format of the data we have to work with, this increases complexity and introduces the non-trivial part of creating a network that we can easily analyze and work with. An extra issue is storing road segments instead of complete roads or connected roads, this way the data in some cases contains "holes" in the geometry, which we have to deal with later on.

To be able to evaluate our results, we also need ground truth data. Directly measured mobility data is rare, it is often conveyed with surveys - it may not be available for all regions or time periods unless the survey is nationwide. In our case, the Hungarian Statistical Office collects census data every now and then, the last collection that we have is from 2016.

1.4. Literature review

Human mobility has been studied extensively throughout multiple centuries, so here we only focus on the works which focus on constructing origin-destination matrices, and other works which are related. Data collection of local location-location commuting has existed for a long time, and researched by local authorities (one such example is Ray S. Keller's work on the statistics of New England's traffic[vii]). From measurements, they derived values often with simple statistical fits. Classically, the task was formulated as a framework for urban traffic planning (UTP), named the "four-step model" from its 4 major steps: determining the proportion of trip endings and/or starts for each region (trip generation), calculate a distribution of origin-destination pairs traffic (trip distribution), divide down proportionally to transportation method (modal split or model choice) and allocate the trips to routes (route assignment or traffic assignment)[viii]. In our case, the last two steps are not relevant (except that we do something similar to the last step as part of constructing one of our models), and step 2 is the most crucial. One of the earliest articles of origin-destination inferring using a model is the 1960 paper of Morris and Burch[ix] which constructs the traffic amounts on traffic lines in Washington, DC, USA. This paper uses the influential method that most researchers in the mid-20th century work use, the gravity model of human mobility, formulated by Stewart[x]. This method calculates mobility based on region population sizes and their distance, generally, the estimate is proportional to the location population numbers, and inversely proportional to some function of the distance (originally first and second power, the second power formula imitates the Newtonian gravity formula, hence the name; lately, power law functions have been used mostly as experimental research indicates it fits the best). This method however doesn't account for varying location functionalities, the method cannot lower bias to population size for situations such as an accounting for less populated areas which attract a lot of commuters, workers, and doesn't capture relationships between locations.

Working with traffic data accounts for these concerns. Models based on traffic were the most common before the popularization of the gravity model around the 1960s. In the 1970s and 1980s, a family of approaches using traffic data emerged. These approaches handle the origin-destination matrix inference as an optimization problem. They make an assumption based on which the set of O-D values that satisfy this assumption are considered. This assumption is that given a matrix of probabilities, which describes the probabilities of crossing a road (segment) upon travelling from origin A to destination B, the solutions should satisfy all equations described by the commuters, where each equation is a road's traffic count (captured from data) equaling the weighted sum of commuters on the road, where the weights are the probabilities. Among the solutions that do satisfy these equations, they choose one that best optimizes a decided objective function – and the models in this category differ in the chosen objective function. Willumsen formulated these methods in this form[xi]. Wilson proposed entropy maximization as objective function[xii] in 1970, which was also put forward later by Van Zuylen and Willumsen[xiii], and used by many other authors. Among these approaches is a model proposed by Bell[xiv] in 1983, which among the feasible solutions selects the one with the highest likelihood of satisfying a given probability distribution. This prior parameter, denoted by $\boldsymbol{q}$ (column vector of J length) is a probability mass function (input space is discrete) over the origin-destination pairs; interpreted as how likely a commuter is going to commute from origin i to destination j, i.e. the proportion of the total commutes that originate from i and terminate in j. We further refer to this as the Bell-model, and it will be the core topic of Chapter 3.3. Throughout the years, (other) statistical methods have been developed, some also build on the same assumption and optimize an objective function, such as the maximum likelihood (assuming Poisson-distributed systems[xv], or normally distributed[xvi]), some are independent from this assumption[xvii] [xviii] (including least-squared error models and iterative proportional fitting). Machine learning (ML) approaches also have been used even for static predictions[xix].

A more similar method to the gravity was published as the radiation model in 2012[xx], which is more universal as can be applied to study specific events (such as migration waves), overcoming a key application issue of the gravity model.

Modern approaches include machine learning approaches and are mostly used in the case of dynamic traffic prediction – thus mostly working with traffic data. In dynamic case, time-series prediction models (statistical, control theoretical e.g. Kálmán-filter, ML-based such as LSTM) can be used[xxi], leveraging űrecent O-D values to predict upcoming ones.

Survey papers and thesis works have collected the works in this field, such as by Mohammed[xxii], whose surveys inspired this literature review.

2. DATA

We obtain data from two different sources, the first one for ground truth comparisons, and the latter one for model construction.

2.1. 2016 Census data from the Hungarian Central Statistical Office

The data includes two datasets:

- Dataset 1: Settlement statistics: population, geographical attributes (latitude,longitude)
- Dataset 2: Commuting numbers from one settlement to another: includes two types of commuting values for each O-D pair: number of work commuters, and school commuters.

This data has already been used by Ódor et al. for describing epidemic seeding effects for the case of the Covid-19 outbreak[xxiii]. It contains every settlement (~3200) in the first dataset, and every O-D pair of settlements that have at least 1 commuter inbetween (~93000 O-D commuting pairs). As this dataset is from 2016 its mobility flow values may not be up to date, which came as a problem for researchers during the Covid crisis. Therefore, this shouldn't be taken as precise ground truth data, but as a benchmark to compare proportions to. At high accuracies, a model with higher ground truth evaluation accuracy may not actually mean a better model in practice (due to the dataset being outdated) – the most important thing is that a good model should be proportionate and correlate well with the given origin-destination flow values.

2.2. 2022 Hungarian National Public Roads cross-sectional traffic data

This is our observational data: it contains the traffic values on national public roads such as highways, main roads and access roads. The data includes the vehicle counts on each road segment (vehicles categories include car, bicycle, heavy vehicles). Road segments are stored with the road's name, measured traffic on the segment and geometry data, representative of the geographical positioning. It also contains some measurement informations, error rates.

The main dataset from this source that we used during the time of this work was the 2022 annual dataset, which contains the most amount of road segments and geographical information: it includes nearly 14000 road segments. Many access roads are represented as one road segment, but for comparison some highways were separated into around 100 segments. We have also obtained access to monthly data from 2016 to 2022 from the Hungarian Public Roads Zrt. that we later will use for inferring mobility flows for each month, analyzing and comparing the results.

3. METHODS

3.1. Overview of methods

We work with two types of previously mentioned model categories: a gravity model, and constrained optimization models.

The original gravity model estimates each O-D pair traffic with a product of the populations divided by the distance, times a constant:

$$T_{ij} = K \frac{P_i \, P_j}{r_{ij}} \ (1)$$

Here, T_{ij} is number of trips from zone i to zone j (what we want to infer), P_i, P_j are population sizes, and K is a constant of proportionality. A more general relation can provide better results:

$$T_{ij} = K \frac{m_i^{\alpha} m_j^{\beta}}{f(d_{ij})} \ (2)$$

Here, m_i, m_j relate to the number of trips leaving location i / attracted by location j, in our case we still use population sizes, but it can help that the number of trips doesn't grow linearly based on population. $f(d_{ij})$ is a function of the distance between zones i and j, named deterrence function, as the number of trips decrease as distance increases.

The constrained optimization models as previously describe, minimize or maximize some objective function subject to an assumption (explaining the need for optimization later):

$$\max f(\boldsymbol{t}) \qquad \text{assuming} \ \ \boldsymbol{v} = \boldsymbol{Pt}$$

where $\boldsymbol{t}$ is a vectorized form of the O-D matrix, it contains the O-D pairs, this is what we infer (J-size column vector), $\boldsymbol{v}$ is the traffic volumes vector, which we get from the traffic measurements data (I-size column vector), and the P-matrix is what connects the two to gather the equations as constraints ($I \times J$ shape matrix). What the P-matrix represents is the likeliness of passing through a certain road, when travelling from origin A to destination B. In order to construct the constraint equations, we need to pre-determine the P-matrix, esentially creating a look-up-table for the p_{ij} probabilities, which theoretically equal to the proportion of the traffic between O-D pair j (e.g. origin A and destination B) that passes through this road. This is practically hard to do and it is not trivial how it can be estimated, it is something that strongly correlates with what we are actually trying to infer. Furthermore, the route selection proportions can change in time, therefore the P-matrix also needs to be changing over time, which is not something we can correct.

A problem that comes up with this formulation is that in practice, in almost all cases the equations have multiple solutions. This is because of cases like three cities laying on one traffic line, such as Budapest-Kecskemét-Szeged, and having to cross the inbetween location when going from one end to the other. For example, in the case of commuters passing by Kecskemét when going to Szeged, we cannot distinguish between cases of one commuter going from Budapest to Szeged without stopping at Kecskemét, and one-one commuter going from Budapest to Kecskemét, and Kecskemét to Szeged respectively, for both cases in a simplified example we obtain the same traffic. Gergely Ódor put it a bit more insightfully: we would like to know the number of commuters for 3 (or 6 if directed) location pairs, but we only have 2 traffic segments to infer from, for our 3 variables, which of course results in an underdetermined system of equations. (There are papers which work with data that tracks vehicles, thus are much more straightforward for estimation[xxiv], however this is not available generally and for us.)

To cope with this, these constrained models typically introduce some objective function among the ones which we have already stated. The constrained optimization model we primarily use is the Bell-model, which maximizes the probability of the O-D vector values being stochastically generated from a given probability distribution. We have experimented with the entropy maximization, and even entropy minimization objective function models, but found that the performances of these models are underwhelming and may not outperform the implemented gravity model. Thus, we only discuss the Bell-model further on.

3.2. Gravity model

In our case, (2) has a power-law deterrence function, so $f(d_{ij}) = {d_{ij}}^{k}$, where k is some positive number. We set $\alpha = 1$, K does not matter as we later divide down / multiply to make the total O-D values sum equal to the ground truth data sum for comparison. The parameters

$\beta,\ k$ are fitted on the ground truth data, this adds some biased prediction power, but this model is only intended for benchmark comparisons with the Bell-model.

We use the scikit-mobility software package provided by Pappalardo and others[xxv].

3.3. Bell-model

The Bell-model uses an objective function, that finds the maximum probability O-D vector being stochastically generated from an a priori probability distribution $\boldsymbol{q}$ (this builds on the idea of Bayesian statistical inference), assuming that both $\boldsymbol{q}$ and $\boldsymbol{t}$ are multinomially distributed (independent trials probability distribution, with multiple categories, a generalization of the binomial distribution.) In this case, the objective function is defined with this formula:

$$\max \mathrm{P}(\boldsymbol{t} \mid \boldsymbol{q}) \rightarrow f(\boldsymbol{t}) = \frac{\left(\sum_j t_j\right)!}{\prod_j (t_j!)} \cdot \prod_j {q_j}^{t_j}$$

This is the model and thus objective function that we will work with, with one minor modification, that we take its logarithm $F(\boldsymbol{t}) = \ln\left(f(\boldsymbol{t})\right)$ to not compute with too large numbers. Using the asymptotic formula $\frac{d\,x!}{d\,x} = \ln(x)$, the partial derivatives can be written in simpler form: $\frac{d\,F(\boldsymbol{t})!}{d\,t_k} = \ln\left(\sum_j t_j\right) - \ln(t_k) + \ln(q_k)$, which is useful when implementing the gradient of the function.

3.3.1. Practical implementation, modifications

We use the scipy optimize function to maximize the function (minimize the function multiplied by -1). It requires an initial estimate to start iterating from. If the function was concave, the starting point would not matter in result, only in computation time. Because our objective function isn’t convex in many domains, the initial guess is important, as the method will likely iterate into a local minimum, and not a global minimum. Hence, a good initial guess is needed (experimentally, we found that small initial guesses tend to converge to a very low average value guess, and with high guesses similarly we get very high average values, therefore not only proportions, but the total average also matters for a good initial guess). We use the gravity model results as input: we see this as an opportunity to show that the Bell-model can improve on the gravity model results if starting from it.

The q probabilities in the article are taken to be a normalization of an initial guess, we do the same with the gravity model. The choice of a heuristic to construct a P-matrix is important. For a starting point, we implemented an all-shortest-paths heuristic to construct the matrix. The shortest paths algorithm works on graphs, later we will see how we turn our data to a network. The algorithm:

- We initialize the P-matrix with zero values
- For all pairs of locations, all shortest paths are computed
- For each of the *N* shortest paths between two locations, $\frac{1}{N}$ is added to stored values of the edges of the path

As an example, if we find only two routes which require three steps to get to the destination from the origin, and there are no routes with less steps, then for each edge of each route, add 0.5 to the respective value of the edge (road segment) in this origin-destination column. If an edge is contained in both routes, its probability will equal to the maximum 1 value. This method is quite primitive and doesn't account for distances or time.

(Note: heuristic P-matrix creation is much related to the route assignment step in the four-step model, so one could derive ideas from the methods for route assignment. Historically, most of the methods are game-theorical and search for traffic equilibrium points (an equilibrium is where commuters are distributed among paths in such a way that nobody can improve their travel time by changing route, which may not be everyone choosing the theoretically shortest time route if we add delays on roads based on traffic. These ideas were not tried out in this work but are interesting for future considerations.)

Regarding the constraints, in the original paper, constraints are taken as necessary to be satisfied. In our case, because of measurement errors and P-matrix inconsistencies, some constraints actually conflict. We do not want to throw away linearly dependent equations, because on a large scale this will result in a lot of deleted constraints, however we cannot keep everything as is, because no O-D values satisfy conflicting values. The solution we found is to dissect the constraints into two categories: linearly independent (from other constraints) ones, and linearly dependent ones. The independent equations are kept as constraints, while the dependent ones are subtracted as loss from the objective function. The choice of loss function is not arbitrary, because $F(\boldsymbol{t})$ grows asymptotically as *O(xlogx)* (derivative grows as *ln(x)*, from the previously mentioned formula), L2 error would be too strong, as error grows faster than the objective function does. L1 error is acceptable, even if it doesn't grow the same order as $F(\boldsymbol{t})$ does, which can cause issues in the coefficient choice, however L1 is not derivable at 0, which is an issue for the optimizer method, therefore a smooth approximation of L1 is used. This is one potential loss function, the other tried „entropic absolute error" loss function: $L(|\boldsymbol{l}|) = \sum_{j \in dep.} |l_j| \ln(|l_j| + 1)$ where $l_j = v_j - \sum_k P_{jk} t_k$ the error from constraints, this loss function grows asymptotically as the objective function does. In practice, both methods perform similar, the optimal coefficient for the second loss method is more stable, but calculations are considerably slower with this loss function.

3.4. Pipeline

A diagram describes the key steps of the process of creating both models and calculating results on Figure 1.

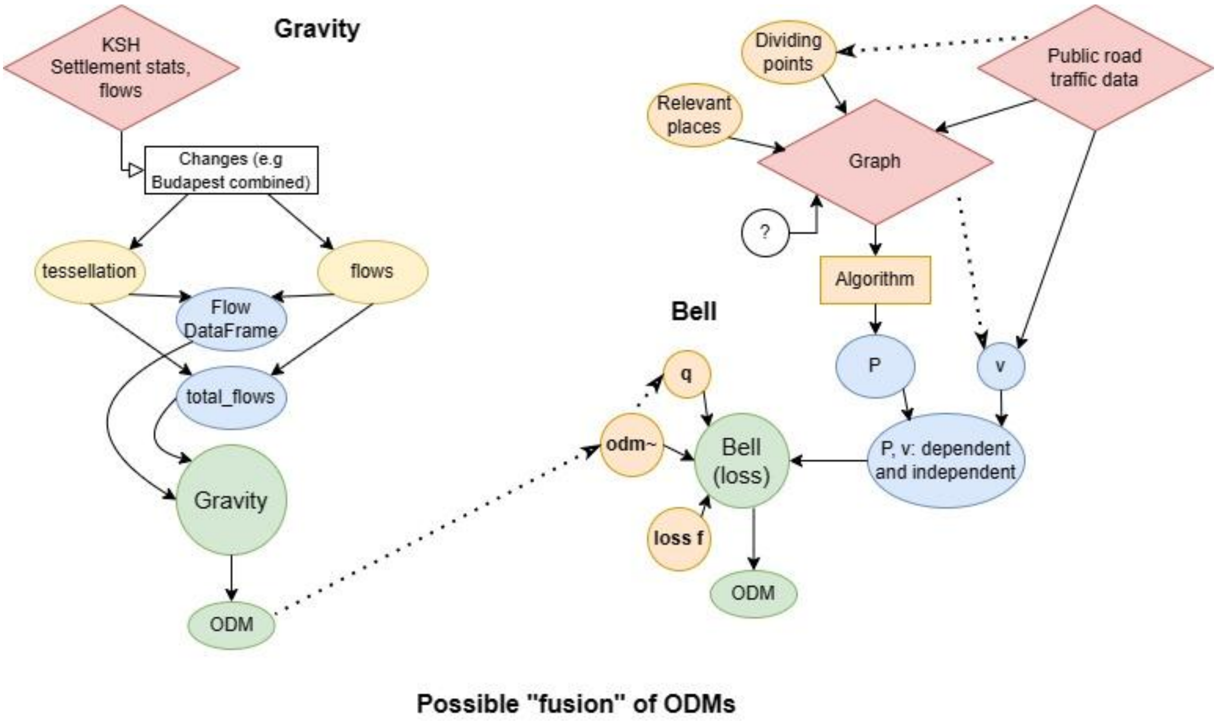

Figure 1: The pipeline of the two models.

The gravity model calculates total flows to each location, storing the location attributes (including total flows) in a tessellation table (dataframe), then fitting a gravity model on the data. The output of this model is reused for the Bell-model. The Bell-model takes the geometric data and creates a network from it, using division points as nodes. The relevant locations (e.g. settlements) are also inputted and placed in the graph. The P-matrix algorithm calculates the P-matrix values, in the next step we construct the constraints and divide them into the dependent and independent constraints categories, and then run the constrained optimization with the lossy objective function.

4. Results

4.1. Proof-of-concept graph

To test out models as quickly as possible, and to immediately see what problems might arise upon building the pipeline, we first created a small, manually crafted graph that shown on Figure 2 that resembles the key traffic hubs based on the graphs seen on Figure 3.

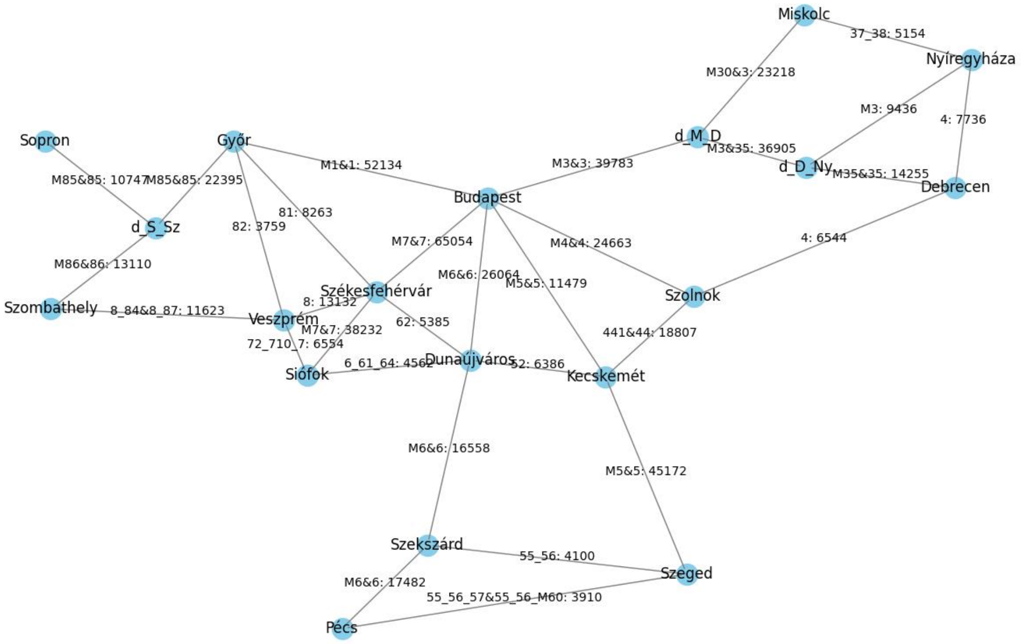

Figure 2: Proof-of-concept small network.

The highest traffic roads were extracted, most common roads between nearby locations were added, and a multigraph was created (two nodes can have multiple edges between them, e.g. road 3 and highway M3). The major locations on these roads were included in the graph, plus three highway division points. Each edge was linked to a set of road segments in the dataset, the traffic value was chosen to be the minimum value in this set. The parallel edges for the computation were combined into one to create the simple graph on Figure 2.

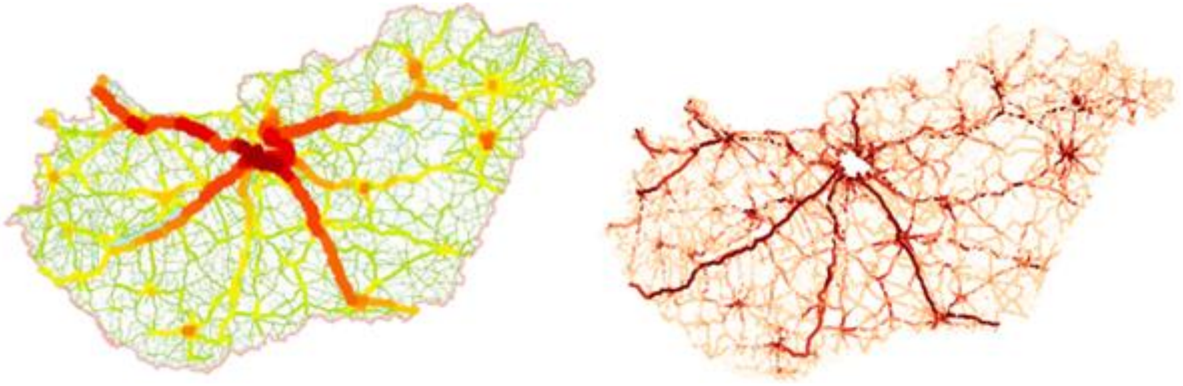

Figure 3: Highest traffic Hungarian national public roads.

The Bell model was run on this small network. The gravity model used the same locations for the O-D matrix, but used the population dataset for inference.

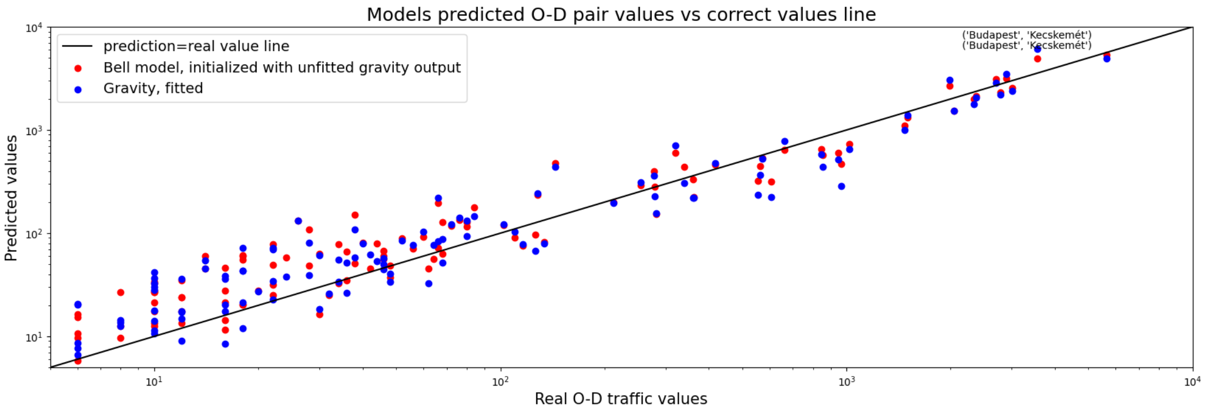

Figure 4: Comparison of the two models: each ground truth O-D value is in increasing order. The real values would appear on the line (plotted as a line for convenience). We see that mainly on the larger values, the Bell model estimates closer to ground truth data, and is more often closer to ground truth

The results of the two models are clearly comparable, as they make estimates on the same origin-destination pairs. We compare the estimates to the ground truth data, and evaluate them.

	Correlation	***MSE***	***MAE***	***Relative difference (sq.)***	***Better prediction***
Gravity	95.8%	62479	~124	**1103**	49
Bell (from Gravity)	**97.4%**	**37298**	**~94**	1138	**71**

As we can see, in almost all benchmarks, the Bell model does better, most notably reducing the mean squared error to almost half of the initial value. Note, that the Bell model uses an unfitted gravity model for base, while the comparison is made to a fitted gravity model. As the MSE decreased, while relative difference increased, this indicates that the model does considerably better on large O-D values, whilst doing not better on small estimates. The amount of better predictions is also substantially higher. This indicates that indeed, the Bell-model seems to work better with good setup, as we expected from the traffic information.

5. Future work

5.1. Scaling up

As we have only worked with a small manual example, we need to scale up the network and test our results on this larger scale too. If the network is extensive enough (e.g. all settlements are tagged for computation that have above 1000 inhabitants), with improvements on the issues stated in 5.1.1, the output result can be used as an origin-destination matrix for Hungary.

5.1.1. Technical difficulties

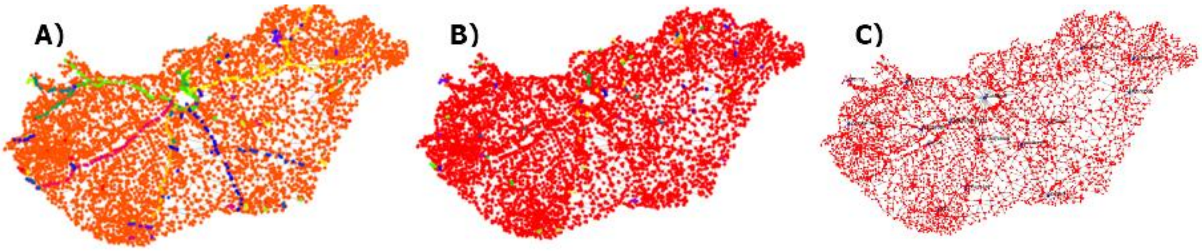

Figure 5: Arising network construction issues on the complete network: A) is the network after just connecting road segments based on endpoints. Highways, major roads appear as separate components, over 100 components in the graph. B) Connecting intersecting roads, common especially between highways and access roads leading to highways, this procedure connects these major roads to the main component. C) The locations are added to the network and nodes around are combined into one location node + component connecting

A major challenge which appears at high scales, that we did not experience with the small network, is the reliable construction of the network. For trustworthy results, we need to run our computations on a connected, simple graph that contains the settlements as separate nodes near their geographical position. As seen on Figure 5, just adding a node for each road segment and connecting the endnodes does not necessarily result in a connected graph. In fact, most highways stay disconnected from the major component of the network. For this reason, intersections between roads are also searched and new edges are being added, which is particularly useful for connecting highways and other main roads to the main component network, as many access roads cross the highways but are not connected in the first case. We then have to add the settlements to the network as new nodes, and within some range, combine them with the nodes appearing in the graph – this range should depend on the settlement size. If there are still multiple components left, we can connect them with simple heuristics.

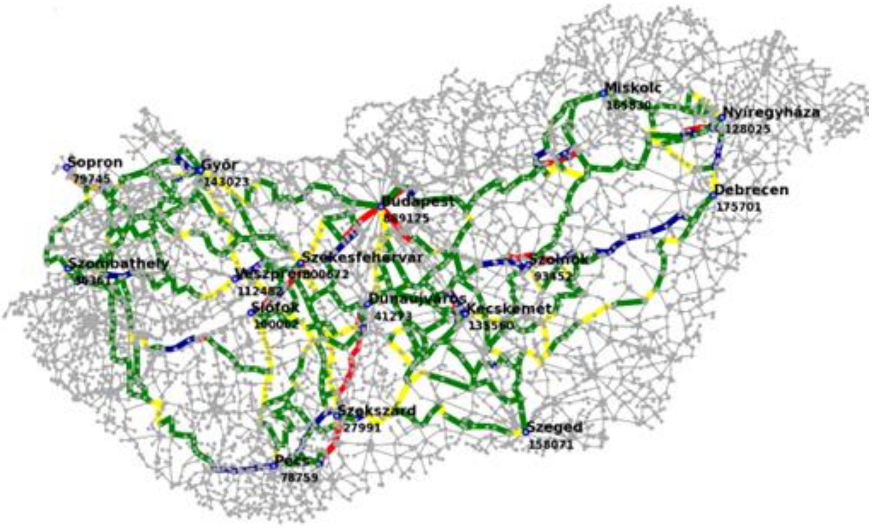

Figure 6: The shortest paths between major locations in the large graph. Red roads represent highways (130 km/h), orange roads near Sopron have speed limit 110 km/h, yellow roads have limit 90 km/h, green roads 70 km/h, and blue roads 50 km/h. The unweighted shortest path algorithm doesn't account for time or distance

Another issue that is more serious on this scale is the P-matrix creation heuristic. As Figure 6 shows, just using a shortest path algorithm based on least edges in the graph does not yield realistic paths on this level. The graph clearly shows that most road segments included with this heuristic have speed level 70, which means they are not among the major road types, and typically have less traffic. We would higher traffic if including the major roads, therefore with this P-matrix the model would underestimate. A proposed, and partially implemented alternative for this algorithm is the k-shortest-times algorithm, which includes those *k* shortest time paths between two locations, but this may be marginally slower.

Computational issues might also arise. Just the calculations and optimization took ~46 seconds for 16 locations. If we include several hundred locations, this can grow very fast. Network creation is also slow, finding intersections took 25 minutes, making the overall procedure 30+ minutes long. It is not clear how procedures can be paralelized, or speeded up.

6. Conclusion

The goal of the project is to build models that use traffic data for the estimation of the origin-destination matrix for Hungarian settlements, on a scale that includes at least the settlements over 1000 number of inhabitants. We created a model and tested it on a simplified network, comparing the results to the gravity model and seeing positive results. We made good progress on generalizing our method to work well on a larger network, meanwhile processing the geographical data of public roads to create such a large network. Next steps are to improve on the shortcomings of our heuristics, and once reliably working on large networks, compare results to other methods, scaling up the analysis, and use the same solutions on the monthly datasets.

[i] Hyungsoo Woo, Okyu Kwon, and Jae-Suk Yang, 'Global Transmission of COVID-19 — A Gravity Model Approach', *International Journal of Modern Physics C* 34, no. 04 (April 2023): 2350055, https://doi.org/10.1142/S0129183123500559.

[ii] MRKV Zainuddin et al., 'Dynamics of Malaysia's Bilateral Export Post Covid-19: A Gravity Model Analysis', *Jurnal Ekonomi Malaysia* 55, no. 1 (2021): 51–69.

[iii] Zixuan Liu and Raphael Stern, 'Quantifying the Traffic Impacts of the COVID-19 Shutdown', *Journal of Transportation Engineering, Part A: Systems* 147, no. 5 (May 2021): 04021014, https://doi.org/10.1061/JTEPBS.0000527.

[iv] Hocheol Lee et al., 'The Relationship between Trends in COVID-19 Prevalence and Traffic Levels in South Korea', *International Journal of Infectious Diseases* 96 (1 July 2020): 399–407, https://doi.org/10.1016/j.ijid.2020.05.031.

[v] Júlia Koltai et al., 'Monitoring Behavioural Responses during Pandemic via Reconstructed Contact Matrices from Online and Representative Surveys', *arXiv Preprint arXiv:2102.09021*, 2021.

[vi] {Citation}

[vii] Philip B. Herr, 'The Regional Impact of Highways' (Massachusetts Institute of Technology, 1959).

[viii] Anders Peterson, 'The Origin-Destination Matrix Estimation Problem: Analysis and Computations' (Institutionen för teknik och naturvetenskap, 2007).

[ix] Morris Robert L. and Burch James S., 'Analyzing and Projecting Travel Data', *Journal of the Highway Division* 86, no. 4 (1 December 1960): 55–58, https://doi.org/10.1061/JHCEA2.0000130.

[x] John Q. Stewart, 'Demographic Gravitation: Evidence and Applications', 1948, https://api.semanticscholar.org/CorpusID:158495727.

[xi] L. G. Willumsen, 'Simplified Transport Models Based on Traffic Counts', *Transportation* 10, no. 3 (September 1981): 257–78, https://doi.org/10.1007/BF00148462.

[xii] Henk J. Van Zuylen and Luis G. Willumsen, 'The Most Likely Trip Matrix Estimated from Traffic Counts', *Transportation Research Part B: Methodological* 14, no. 3 (1 September 1980): 281–93, https://doi.org/10.1016/0191-2615(80)90008-9.

[xiii] Van Zuylen and Willumsen.

[xiv] Michael GH Bell, 'The Estimation of an Origin-Destination Matrix from Traffic Counts', *Transportation Science* 17, no. 2 (1983): 198–217.

[xv] M. E. Ben-Akiva, 'Methods to Combine Different Data Sources and Estimate Origin-Destination Matrices', *Transportation and Traffic Theory*, 1987, https://trid.trb.org/View/315716.

[xvi] Henk J. van Zuylen and David M. Branston, 'Consistent Link Flow Estimation from Counts', *Transportation Research Part B: Methodological* 16, no. 6 (1 December 1982): 473–76, https://doi.org/10.1016/0191-2615(82)90006-6.

[xvii] Anselmo Ramalho Pitombeira Neto, Francisco Moraes Oliveira Neto, and Carlos Felipe Grangeiro Loureiro, ‘Statistical Models for the Estimation of the Origin-Destination Matrix from Traffic Counts’, *TRANSPORTES* 25, no. 4 (30 December 2017): 1, https://doi.org/10.14295/transportes.v25i4.1344.

[xviii] Adom Giffin and Ariel Caticha, ‘Updating Probabilities with Data and Moments’, in *AIP Conference Proceedings,* vol. 954 (American Institute of Physics, 2007), 74–84, https://pubs.aip.org/aip/acp/article-abstract/954/1/74/840318.

[xix] Behrang Assemi et al., ‘Improving Alighting Stop Inference Accuracy in the Trip Chaining Method Using Neural Networks’, *Public Transport* 12, no. 1 (2020): 89–121.

[xx] Filippo Simini et al., ‘A Universal Model for Mobility and Migration Patterns’, *Nature* 484, no. 7392 (2012): 96–100.

[xxi] Xi Xiong et al., ‘Dynamic Origin–Destination Matrix Prediction with Line Graph Neural Networks and Kalman Filter’, *Transportation Research Record: Journal of the Transportation Research Board* 2674, no. 8 (August 2020): 491–503, https://doi.org/10.1177/0361198120919399.

[xxii] Mohammed Mohammed and Jimi Oke, ‘Origin-Destination Inference in Public Transportation Systems: A Comprehensive Review’, *International Journal of Transportation Science and Technology* 12, no. 1 (2023): 315–28.

[xxiii] Gergely Ódor et al., ‘Switchover Phenomenon Induced by Epidemic Seeding on Geometric Networks’, *Proceedings of the National Academy of Sciences* 118, no. 41 (12 October 2021): e2112607118, https://doi.org/10.1073/pnas.2112607118.

[xxiv] Nanne J. Van Der Zijpp, ‘Dynamic Origin-Destination Matrix Estimation from Traffic Counts and Automated Vehicle Identification Data’, *Transportation Research Record* 1607, no. 1 (1 January 1997): 87–94, https://doi.org/10.3141/1607-13.

[xxv] Luca Pappalardo et al., ‘Scikit-Mobility: A Python Library for the Analysis, Generation and Risk Assessment of Mobility Data’ (arXiv, 4 June 2021), http://arxiv.org/abs/1907.07062.