Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2019) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

The 10th International Conference on Ambient Systems, Networks and Technologies (ANT)
April 29 – May 2, 2019, Leuven, Belgium

An Automated Gradual Zoning Approach for large-scale Transport Models

Falko Nordenholz[a,]*, Simon Metzler[a], Christian Winkler[a]

[a]*Institute of Transport Research, German Aerospace Centre, Rutherfordstraße 2, 12489 Berlin, Germany*

Abstract

Along with the advance of information technologies and open data availability, transport models have become more complex in the last years. However, the basic structure is often neglected in model design. Transport analysis zones which form the spatial base of a model are often drawn alongside administrative boundaries and remain untouched, while the model evolves. This paper presents an automated approach to design traffic analysis zones as first step of a model, taking into account the administrative structure and possible data availability dimensions. A quality control step is included into the model framework.

Keywords: Transport model; zoning algorithm; MAUP; Geographical Information Systems; TAZ; open data

1. Background and Motivation

Most transport models are based on a spatial subdivision of the area of investigation. The result of the subdivision process is a zoning system that subdivides the study area into smaller units. These units are named traffic analysis zones (TAZ) and are used to represent the corresponding land-use data such as number of inhabitants, employees, workplaces, and so on. Moreover, in the context of a transport model any address within the area can be allocated to a TAZ. The number of defined TAZ, and therefore the level of spatial detail, depends strongly on the study area size and the projects to be evaluated on the basis of the model results. However, available budgets of time and money for the model development are also crucial and have to be met. There are some recommendations in the literature for

* Corresponding author. Tel.: +49 30 67055 599.
E-mail address: falko.nordenholz@dlr.de

defining appropriate zoning systems. Most of them are based on experiences and have a focus on transport planning and modelling aspects. Corresponding to this, zones should be homogeneous in their composition regarding land-use and/or population composition and be of similar dimensions in travel time units and therefore be smaller of size in dense areas and larger in rural areas [1].

Besides modelling aspects also the availability of spatially differentiated land use data as a major input for transport models is most important. These data sets are typically available at administrative levels as counties, municipalities or higher order but rarely more detailed. To fulfil modelling needs often a higher level of detail is necessary. Therefore, in most cases administrative zoning structures are used and then further spilt into smaller zones that represent the final TAZ. This is a costly but, in particular for urban or regional models, necessary process. For large-scale models as for whole countries such an approach is generally not feasible. The use of administrative zoning systems as TAZ is reasonable when only long-distance trips are of interest. However, when also short-distance trips should be considered further differentiation is necessary.

A solution to the conflict between the necessity of a high number of TAZ on the one hand and the related high effort on the other hand is not straightforward but a new methodology and emerging data sets based on a raster system provide promising opportunities to overcome the problem. The new method is a gradual approach to develop a zoning system on the basis of an attributed geographical raster grid with population information. The concept enables to generate zoning systems with different zone sizes and was, to our knowledge, firstly used by Moeckel and Donelly for the state of Georgia in the USA [2]. Molloy and Moeckel (2016) [3] extended their work and developed the pyGr framework as a configurable Python software package. An advantage of raster data systems is that they provide data of a higher level of spatial detail at equally-sized zones, limited only by privacy standards when needed. Furthermore, the system is independent of administrative structures, which are continuously changing.

The German national transport model DEMO [4] is a large-scale transport model with over 6,500 TAZ that has been developed and applied at the Institute of Transport Research for many years. The zoning system was defined in 2010 and is based on administrative structures as of 2008. It was further differentiated within urban areas on the basis of a zonal system provided by a commercial data supplier. However, since then many administrative boundaries have been changed and data updates are getting more and more complicated. Furthermore, model results suggest that a higher level of detail is needed in particular within urban areas. For these reasons a new zoning system for DEMO needs to be defined. The traditional approach starting with administrative zoning systems and a gradual differentiation by hand is not practicable in light of the model dimension and an alternative approach has to be developed.

In this paper we present an automated approach to designing TAZ, taking into account both, the administrative structure and possible data availability issues. For this approach we are using new raster data sets provided by the German administration to subdivide administrative zones. As open-access data sets have become more useful, we are only considering openly available data as input for DEMO. We use the pyGr framework as the methodical basis and extend the concept by a post-processing procedure and an additional quality control algorithm. The resulting approach should be adaptable to different input data formats and spatial entities even at a large scale.

2. Data

Geographic data for transport modelling purposes are available in two dimensions: Firstly, there are polygon data: Examples are “population of a municipality” or “number of employed persons in a county”. These data are attributes of the containing polygon. Secondly there are point data, e. g school addresses with an attribute “number of students on site”. The goal of a zonal structure in a transport model is to aggregate these different data sources to one single aggregation level, i.e. TAZ. It can be seen, that point data can easily be aggregated to any desired aggregation level, while polygon data are often only provided for existing administrative structures. However, administrative structures often do not fulfil all desired requirements of TAZ. Moreover, they are usually more or less arbitrary, as administrative boundaries are subject to change. These aspects speak against using administrative structures as TAZ, even though they are most important land-use data sources. Regular changes in these official structures, however, cause necessary modifications to distribute land-use data to the model TAZ, which is an

unwelcome and possibly effortful effect. In Germany, due to administrative alterations, the number of counties decreased by about 25 percent from 543 since 1993 to 401 in 2018 [5].

An independent and spatially highly differentiated alternative is a raster system, which might overcome many shortcomings of the system based on administrative boundaries. The German administration has recently implemented such a raster zone system, which splits the country into equally-sized zones (e.g. 1 km × 1 km), ignoring any administrative or natural boundary within the zone. The zones follow specifications according to the INSPIRE technical guidelines [6] on geographical grids and are basically available for the entire EU. Data from the (latest) 2011 census are available at this zonal level; however the annual update is carried out at administrative level only. Therefore, any newly implemented system has to integrate the existent administrative boundaries. Hence, it is necessary to combine three different aggregation levels: Point data (not aggregated), Grid raster data and data aggregated at administrative regions. All of these data sets contain information relevant for transport models. However, the latter two cannot be matched without imputation. Data sets at higher administrative levels have to be disaggregated to the smaller raster grid zones by certain rules. The easiest method is to divide the total number of an attribute by the number of raster zones. However this method adds the assumption of equal distribution to the model. Land-use data existing at a smaller scale in turn can help to solve this issue: Workplaces can be more precisely allocated into industrial areas and residential zones will probably host more population than agricultural areas.

The acquisition of detailed land-use data is a crucial issue for defining TAZ. There are several sources where data can be obtained from. In particular, most relevant are commercial data providers, official data sources and the Openstreetmap project (OSM) [7]. For the development of a new zoning system of DEMO we are solely considering freely available data sets, i.e. official data and OSM data. The latter data source has been chosen because of its open source licensing philosophy and its worldwide availability, as it allows applying the method outside Germany or the EU. However, the known limitations of crowd-sourcing data should be kept in mind and be regarded carefully especially in smaller study areas and at a higher resolution. If required, official land use data could also be used, if it distinguishes at least between residential, industrial, commercial, and retail areas.

3. Model Framework

The high effort of designing a zonal structure for large-scale models makes it feasible to implement an automated process to spare the modeller extensive handwork. Molloy and Moeckel's pyGr software package [8] combines raster data, and administrative boundaries in an iterative process. Their algorithm aggregates population and the number of jobs in a zone. In our work, we call the sum of both values the zone weight. In order to find an optimal solution, the upper and lower thresholds for acceptable zone weights are adjusted within each iteration step. Where necessary, raster data are aggregated by administrative entities. Raster grids are being kept, if their population or number of employees exceeds a predefined number, in order to generate zones at a similar size. They used the framework to design a set of zones within the Munich metropolitan area.

We adapted this approach to apply the toolbox in order to generate a zone structure for entire Germany, using a 1 km × 1km-grid, in contrast to the 100m × 100m-grid chosen by Molloy and Moeckel. We also limited the input data to open data sources. The model requires some basic data specifications. The zone weight is the decision basis, whether a zone is being formed or not. The pyGr framework requires an upper threshold for the desired zone weight and a desired number of resulting zones. We set the zone weight threshold to a low range which leads to a large number of generated zones in the start solution. This increases the degrees of freedom for the subsequent steps. Hence the lower threshold was set to 12,000 jobs and inhabitants and the upper threshold to 15,000. This configuration leads to a number of 19,000 zones as an output from pyGr.

3.1. Model execution

To conduct the zoning in the second stage of pyGr, firstly raster datasets containing the actual zone weights are prepared. For population the German Census 2011 provides data on 1-km-INSPIRE grid level [9] (Fig.1 (a)). Employment data is only available at municipality level. In order to create the raster grid for employment data, a combination of employment figures at municipality level and a landuse data raster is carried out. Land use data is used to distribute the number of employees within municipality borders (Fig.1 (b) and (c)).

Fig. 1. (a) Population in 1 km INSPIRE Grid; (b) OSM-LU-Data; (c) Employment in 1 km INSPIRE Grid; (d) Population + Employment (zone weight) in 1 km INSPIRE Grid

The model computes employment data by raster zone using different weights to distribute the known number of employees in a municipality. The weights are based upon the land use area data and can be configured in the model. We applied the following weights, derived from the original source:

Table 1: Workforce distribution by OSM Land-use type

OSM Land use	weight
Residential	0.2
Industrial	0.55
Commercial	0.2
Retail	0.05

This means 20 percent of a municipality's total employees are allocated in residential areas, 55 percent in industrial areas etc., summing up to 100 percent. This approach generates a distribution similar to the known population distribution. Hence the required zone weights are generated by adding the known population number and the disaggregated workforce (Fig. 1(d)).

3.2. Zoning

The second stage of pyGr consists of two steps. Firstly a quadtree algorithm is used to create quadratic grid zones according to the given weights from the raster data (Fig. 2(b)). Secondly, these zones are split along the given municipality borders (Fig. 2(a)) and merged within the border until there is no neighbouring zone left to merge without exceeding the upper threshold of weights for each zone. This process is repeated until the desired number of zones is created. As a result, municipalities with low zone weights remain one zone of their own, while municipalities with higher weights are split along the grid raster boundaries (Fig. 2(c and d)).

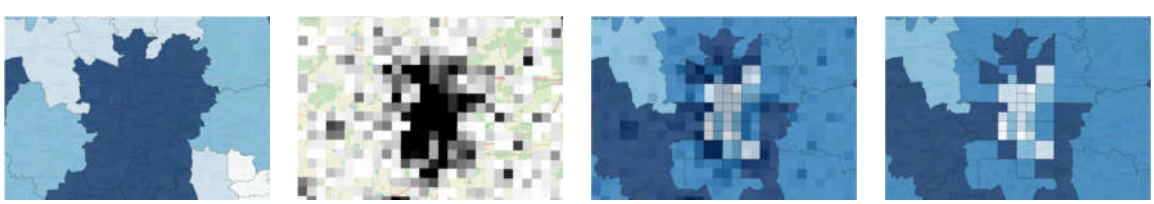


Fig. 2. (a) Municipality borders; (b) Zone weight in 1 km INSPIRE Grid; (c) Zone weight in 1 km INSPIRE Grid and Split Zones; (d) Split Zones

4. Post processing to merge zones

In their original paper, Molloy and Moeckel focussed on the Munich area and used specific regional input data [3]. For a Germany-wide transport model much more input data are necessary, and also further requirements regarding the automated zoning approach apply. Therefore, the model had to be altered and some further processing

is required to adopt it to be suitable for our large scale model. Two administrative layers are involved: municipalities and counties. Enclaves and holes in zones occur and have to be handled.

The PyQGIS-environment was used to tackle these issues based on predefined criteria. For the purpose of the DEMO model, we found an amount of 20,000 jobs and inhabitants per zone feasible. Given, that the sum of jobs and inhabitants in Germany is roughly 112 million (80 million inhabitants and 32 million jobs), this value eventually leads to a number of about 7,000 zones for the study area under the described constraints. This volume is reasonable under the DEMO model specifications, but can be altered in accordance with other models' needs. Therefore, a zone weight of 20,000 was used as target value in the iterative process. We aimed at final zone weights within a range of 0.75 × target value and 2 × target value under a normal distribution. To ensure a reasonable modelling process in the later model steps, they should feature a compact geometry. This will be checked during the process of zone generation in the quality control step explained below.

The zoning is carried out under a defined set of rules. In addition to the criteria from pyGr, we added some rules that apply to larger-scale models rather than to regional models. In no case, zones may exceed county borders, as many input datasets are only available at county level, which would lead to aggregation errors. Municipalities may only be merged as a whole with neighbouring municipalities. Low zone weights are handled by merging small municipalities.

Another qualitative criterion is regarded as mentioned above: For data availability purposes, both zones must have the smallest administrative entity in common. This means, that within a split municipality a merging procedure may only be performed within the municipality. When entire municipalities are merged, both zones have to be within the same county. In cases where a municipality consists of multiple parts (which are not uncommon in Germany) also merges with neighbours of the smaller part alone are allowed. Subsequently, we defined the following set of rules for merging zones:

All of the following criteria have to be met:

- Both zones are within the same smallest administrative entity (municipality or county).
- At least one zone weight is below the lower bound, which is the minimum value to form a TAZ on its own.
- Sum of the zone weights of both potential merging zones does not exceed the desired upper bound, which describes the highest feasible weight for one zone.

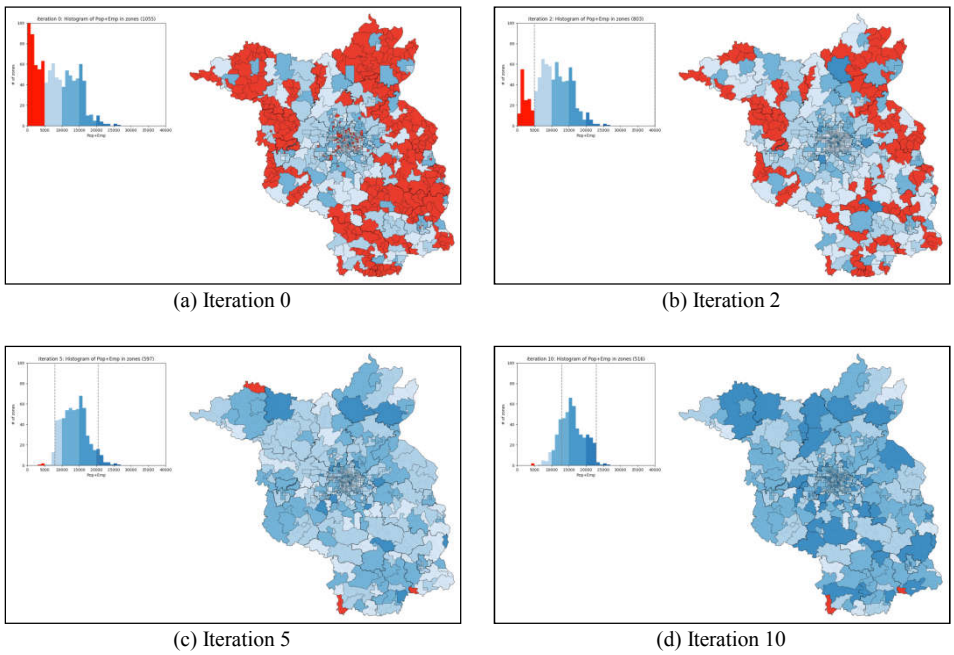
Additionally, one of the following criteria has to be met:

- The border length ratio is bigger than 20%. Border length ratio is defined as common border length divided by the smaller of both zones perimeters.
- One zone weight is below the target value divided by 5 AND border length ratio is bigger than 10%.
- One zone weight is below the target value divided by 200 AND border length ratio is bigger than 3%.
- There is only one neighbouring zone within the same municipality.

If these requirements are met, a merge is conducted. If the requirements are met for more than one relation, those zones with the biggest border length ratio will be merged. The procedure is conducted iteratively. The lower and upper bounds for merging are increased within the process. As a start solution, we define a rather low value for the lower bound to prioritize the merge of low weight zones specifically. The values for the lower and upper bounds are derived from the given target value and evolve during process (see histogram in Fig 3):

- Lower bound = target value × 0.1 + target value × 0.02 × [iteration no.]
- Upper bound = target value × 0.9 + target value × 0.02 × [iteration no.]

For each zone, only one merge-split-operation may be performed per iteration. Therefore, the algorithm loops over every possible zone and over its neighbours respectively to find the neighbouring zone with the best suitable properties. The used quantitative properties are zone weight and the ratio of common border to zone perimeter.



(a) Iteration 0 (b) Iteration 2

(c) Iteration 5 (d) Iteration 10

Fig. 3. Monitoring of zone merging and zone weight distribution during process

Figure 3 illustrates the iterative process. It shows the results of subsequent iterations for the states of Brandenburg and Berlin. The plot beside each map shows the distribution of zone weights. Zones with a zone weight lower than 5,000 are marked in red. Fig. 3(a) shows the state of the zoning after the pyGr process, before the first iteration of the post-processing. There are many small rural municipalities marked in red. These remained untouched from pyGr as their zone weight is low. During the process, the small zones are iteratively merged with neighbouring zones, so that higher attribute values are achieved. By the 2nd iteration (see Fig. 3 (b)) many zones with low weights have been merged. By the 5th iteration (see Fig. 3 (c)) most of these low weighted zones have been removed. During the process, lower and upper bounds increase as described before. This increase is currently fixed within the program configuration, but can be modified to meet special requirements, e. g to meet a certain deviation from the given target value pointed by the grey vertical lines in the histogram (see Fig 3)). It makes sure the zones weights increase continuously while at the same time the distribution function approaches a normal distribution. The final number of zones depends on the desired target value of the zone weight and subsequently the number of iterations.

5. Quality control

The use of an automated approach to define TAZ requires additional quality checks to guarantee reasonable structures. In particular, for large-scale models it would be very costly to check all defined zones individually. Therefore, at the end of the zoning process an automated checking of predefined criteria is integrated into the model algorithm. As already discussed there are several experience-based and recommended criteria for reasonable zones. However, it is not possible to check these criteria automatically and other tests need to be found.

Zoning control mechanisms are often originated in political science: They are commonly used to measure gerrymandering when restructuring electoral districts [10]. For checking compactness we chose the tests suggested by Polsby-Popper (1991) [10]. and Reock (1961) [11].

The Polsby-Popper index is defined as the ratio of the area of the zone (A_Z) to the area of a circle whose circumference is equal to the perimeter of the zone (P_Z):

$$PP = 4\pi \times \frac{A_Z}{P_Z^2} \qquad (1)$$

The Reock index is calculated by dividing the area of the zone (A_Z) by the area of the smallest possible circumscribing circle of the zone (A_{MBC}):

$$R = \frac{A_Z}{A_{MBC}} \quad (2)$$

Zones with higher values are considered to be more compact than those with lower values. Both tests return the maximum value of 1 for a circle. Figure 4 shows examples for different zone shapes with both resulting indices. Administrative borders are often derived from natural conditions such as rivers, mountains or historical matters. This leads to uneven borders, which lengthen an area's circumference. These would result in a low Polsby-Popper index, however the zone can still be considered rather compact if the Reock test performs well (see Fig. 4 (a)). In contrast to this, the rectangle shown in Fig. 4 (b) returns a high Polsby-Popper score, but a low Reock score. The algorithm of pyGr creates squares. A square zone would result in a Polsby-Popper score of 0.79 and a Reock score of 0.64, which are very high values for both indices (Fig. 4 (c)).

Fig 4. (a) Higher Reock score; (b) Higher Polsby-Popper score; (c) Square zone

For our ongoing tests a threshold of 0.15 for desired zones is set for both Polsby-Popper and Reock values. If resulting zones are below this value, they have to be redesigned. However, further tests have to be carried out to test what value is reasonable for defining TAZ. Other issues may arise for "doughnut"- or "island"- shaped administrative regions. Therefore, the integrated algorithm has to be further tested and assessed.

6. Conclusion and Outlook

The presented approach creates TAZ based on population and employment and ensures a homogenous distribution of these attributes. Furthermore it attempts to keep the resulting zones compact and checks the grade of compactness by automated measures. The resulting zones are a mix of administrative borders, which are assumed to be derived from natural conditions and the 1 km × 1 km-raster grid to obtain smaller zones within urban areas. As employment data is not available at this disaggregate level, land use data is applied to distribute a municipality's workforce. The algorithm has been applied for the German national transport model DEMO with a reasonable runtime and is also transferable to other regions.

However, there are still some issues to solve. In general it has to be noted, that the model outcome is a result of the used input data and the configuration. In particular the land use data for the workplace distribution is a known uncertainty, which might be improved by further open data sources. A technical aspect is that a better equilibrium could be achieved, if the algorithm could automatically restructure the result. However, computing efforts and runtime would increase in this case. Further work is also required on the merging process to generate more compact and more homogenous zones. It might also be reasonable to add additional constraints that are more related to travel behaviour, as e.g. a distinction between area types (urban, suburban or rural) during the zoning process. Finally, the quality control step could be improved by a more sophisticated feedback process. It would be more efficient to integrate this step into the merging process, as the check results could directly affect the zoning.

Acknowledgements

The authors gratefully acknowledge the financial founding by the German Federal Government and the Helmholtz-Gesellschaft as part of the project “Transport and Climate (TraK)”. We also would like to thank our colleague Tudor Mocanu for profound discussions and helpful support.

Bibliography

[1] Ortuzar, Juan de Dios, and Luis G. Willumsen. (2011) “*Modelling Transport*”, Chichester, John Wiley and Sons,

[2] Moeckel, Rolf, and Rick Donnelly. (2014). “Gradual Rasterization: Redefining the spatial resolution in transport modelling.”93rd Annual Meeting of the Transportation Research Board, Washington DC

[3] Molloy, Joseph, and Rolf Moeckel. (2016) “Automated design of gradual zones.” *Open Geospatial Data, Software and Standards* **2 (19)**: 19-29.

[4] Winkler, Christian, and Tudor Mocanu. (2017). “Methodology and application of a German National Passenger Transport Model for Future Transport Scenarios.”European Transport Conference 2017, Barcelona

[5] Federal Institute for Research on Building Urban Affairs and Spatial Development. (2015) Laufende Raumbeobachtung - Raumabgrenzungen. URL: https://www.bbsr.bund.de/BBSR/DE/Raumbeobachtung/Raumabgrenzungen/Kreise_Kreisregionen/kreise_node.html (accessed: Nov 7, 2018)

[6] INSPIRE Thematic Working Group Coordinate Reference Systems & Geographical Grid Systems. (2014) D2.8.I.2 Data Specification on Geographical Grid Systems – Technical Guidelines URL: https://inspire.ec.europa.eu/id/document/tg/gg (accessed: Nov 7, 2018)

[7] OpenStreetMap contributors. (2018) Planet dump retrieved from https://planet.osm.org. URL: https://www.openstreetmap.org (accessed: Nov 7, 2018)

[8] Molloy, Joseph. (2017) “pyGr - An automated transport zoning tool.” *GitHub repository*: https://github.com/msmobility/silo_zoneSystem.

[9] Federal Statistical Office. (2016) Bevölkerungsstand nach dem Zensus von 2011 - Gitterzellenbasierte Ergebnisse. URL: https://www.zensus2011.de/SharedDocs/Aktuelles/Ergebnisse/DemografischeGrunddaten.html (accessed: Nov 7, 2018)

[10] Polsby, Daniel D., and Robert D. Popper. (1991) “The third Criterion: Compactness as a Procedural Safeguard Against Partisan Gerrymandering.” *Yale Law and Policy Review* **9 (2)**: 301-353.

[11] Reock, Ernest C. (1961) “Measuring compactness as a Requirement of Legislative Apportionment.” *Midwest Journal of Political Science* **5 (70)**: 70-74.