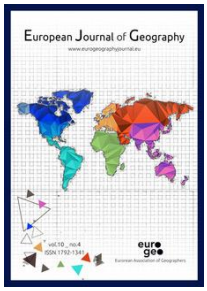

European Journal of Geography

Volume 11, Issue 4, pp. 006 - 021

Article Info:
Accepted: 16/12/2020

Corresponding Author: *k.cheliotis@ucl.ac.uk

https://doi.org/10.48088/ejg.k.che.11.4.06.21

Special Issue: *"Selected Papers from 12th International Conference of the Hellenic Geographical Society (ICHGS): Innovative Geographies II, 2019"*

Using synthetic data for the dissemination of computational geospatial models

Kostas **CHELIOTIS**[1*]

[1] *University College London, Centre for Advanced Spatial Analysis, London, UK*

Keywords:
Datasets,
Synthetic data,
Modelling best practices,
Model communication,
ICHGS-2019

Abstract

Detailed datasets of real-world systems are becoming more and more available, accompanied by a similar increased use in research. However, datasets are often provided to researchers with restrictions regarding their publication. This poses a major limitation for the dissemination of computational tools, whose comprehension often requires the availability of the detailed dataset around which the tool was built. This paper discusses the potential of synthetic datasets for circumventing such limitations, as it is often the data content itself that is proprietary, rather than the dataset schema. Therefore, new data can be generated that conform to the schema, and may then be distributed freely alongside the relevant models, allowing other researchers to explore tools in action to their full extent. This paper presents the process of creating synthetic geospatial data within the scope of a research project which relied on real-world data, originally captured through close collaboration with industry partners.

The publication of the European Journal of Geography (EJG) is based on the European Association of Geographers' goal to make European Geography a worldwide reference and standard. Thus, the scope of the EJG is to publish original and innovative papers that will substantially improve, in a theoretical, conceptual or empirical way the quality of research, learning, teaching and applying geography, as well as in promoting the significance of geography as a discipline. Submissions are encouraged to have a European dimension. The European Journal of Geography is a peer-reviewed open access journal and is published quarterly.

1. INTRODUCTION

The initial spark for this paper came during The Society for Modelling & Simulation International's 2019 Spring Simulation Conference, in Tucson, Arizona, in April 2019. During a panel discussion on best practices in developing artificial societies, the topic of discussion focussed on better ways of publishing and communicating agent-based models (ABMs) and simulations, with a mind towards comprehension and reproducibility by other members of the modelling community. Two main threads emerged during the panel discussion: first, that a standardized approach to writing about and presenting ABMs should be adopted by members of the community; and second, that modellers should ensure that their models are published in such a way as to allow for their replication by others in the community.

Regarding the first point, that of standardization, it has been observed (Angus & Hassani-Mahmooei, 2015) that relatively new computational approaches (such as ABMs for example) often are presented on an ad-hoc basis, with each new piece of work documenting different aspects of a model, leading to confusion and hindering reproducibility. Most of these issues are often overcome with time, when a particular methodology has matured enough and best practices emerge. However, it is often the case that having a standardized approach to writing and communicating a model can speed up the process as well, as adopting a standard ensures that readers are familiar with the structure of an ABM paper, can quickly and at a glance understand how the model works and how to implement and replicate a proposed model. This has been largely addressed by the introduction of ABM protocols, particularly the ODD (Overview, Design Concepts, Details) protocol (Grimm et al., 2006, 2010).

The second point, on model reproducibility, follows closely from the previous one: after establishing a common language for communicating computational models, it is important to make sure that all necessary elements for reproducing a study/model are also provided, which may include for example the reasoning behind design decisions, the code base, and any data that may be required by the model to run. It is this last element, the data, which in recent years has become more and more prominent, and is the main focus of this paper. Kitchin (2014) notes that we are currently seeing a "data revolution", in which more and more data are becoming available, with increasing fidelity and timely publication. However, it is sometimes the case that data is locked behind non-disclosure limitations, where a proprietary dataset may be shared with researchers for the purposes of a study, but researchers are prohibited from further sharing the dataset with the public. This results in an interesting and somewhat worrying point, where new datasets may be used to develop new techniques and methodologies, thus advancing knowledge in a field, but the datasets themselves may be prohibited from being shared, therefore limiting the communication and wider adoption of the derived techniques.

This point on data availability therefore poses a not-insignificant limitation to the reproducibility of computational models that require input data. To resolve such cases, this paper discusses a potential solution in the form of *synthetic* datasets: Made-up datasets to be used as 'stand-ins' for original, un-publishable datasets, that do not reflect any aspect of the real-world, and are therefore free of any disclosure limitations, but are generated in such a way so as to 'look' and 'behave' exactly like the original.

2. CONTEXT

Within the contemporary landscape of data availability, there are multiple classifications on data access, defining the various entities that may access different datasets based on sensitivity and ownership. The Open Data Institute has produced one such classification, termed 'The Data Spectrum'[1] (**Figure 1**), which illustrates why such access limitations might exist: any data more restricted than 'Public Access' may contain sensitive information, individually identifying records, or business-sensitive information. As such, while the sharing of such data for research purposes is welcome by all parties, publication of the dataset itself is highly restricted and often for good reason, thus resulting in researchers working with what is termed here 'un-publishable' datasets. Although the above-mentioned limitation is strongly enforced and respected, research work with 'un-publishable' datasets is quite common, and furthermore researchers continuously publish work based on such data, as there are a number of ways to circumvent such limitations.

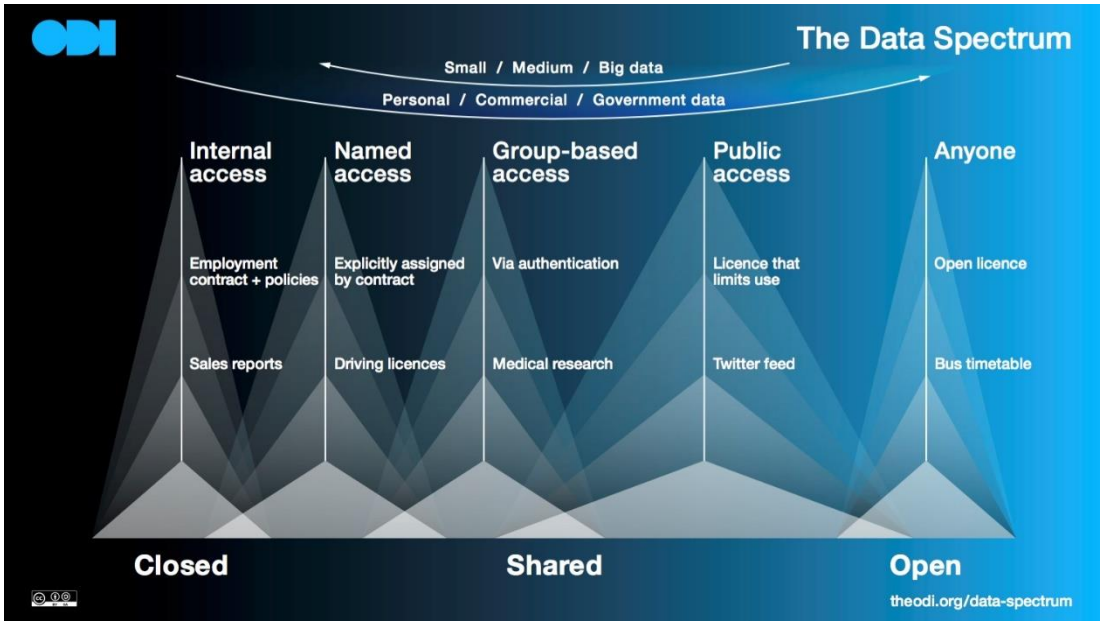

Figure 1. The Data Spectrum (CC-BY license, Open Data Institute)

Within this datascape, by far the easiest way to circumvent limitations of un-publishable data is by avoiding them altogether, for example by working with open data, which by definition has much less restrictive conditions (Kitchin, 2014). Working with open data is regarded as providing benefits both for the research community by supporting better scientific practices and promoting data standards, as seen for example in (Bartha & Kocsis, 2011), as well as at the political level, where open data is regarded as "*providing greater returns from the public investment in research*" (Pampel & Dallmeier-Tiessen, 2014). One such case of using open data in the geospatial research community is the use of OpenStreetMap (OSM), a crowdsourced web-mapping platform (Haklay & Weber, 2008) that has seen increased use in recent years, with research associated with it both using OSM data as input as well as feeding its output back into the community, as seen for example in humanitarian mapping during emergency

[1] https://theodi.org/about-the-odi/the-data-spectrum/, accessed 30/10/2019

response (Dittus et al., 2017). However, it may often be the case that a particular dataset is only available under disclosure-restrictive limitations, and therefore a piece of work can only be generated using 'un-publishable' data. Even in such cases however, examples of published work are plentiful, and highlight two distinct approaches to working with 'un-publishable' data.

One approach to working with 'un-publishable' data is publishing a summary of the data in the form of analysis results and findings. Examples in this approach can be found in analytical work on urban mobility using mobile phone network data (Calabrese et al., 2011; Diao et al., 2015; Reades et al., 2007; Reades et al., 2009) that often makes use of aggregated mobile phone records of individuals, and is therefore highly sensitive information; examining taxi cab driver behaviours using trackers (Liu et al., 2010), which utilizes positional and temporal tracking of individuals, constituting highly sensitive data due to personal identification risks; and analysing outdoor physical activity using mobile electro-encephalography (Aspinall et al., 2013), which utilizes sensitive personal medical data. In all these cases, the underlying data is highly sensitive and therefore never published, but analysis findings summarizing the data in interesting and valuable ways are shared to the public.

The second approach in working with 'un-publishable' data, which has been mentioned already, is the use of synthetic datasets. In cases where an original dataset may not be published, a secondary dataset can be created so that it 'looks' and 'behaves' like the original, and given that it does not represent real-world data (and therefore is not bound by any proprietary or uniquely identifying limitations), it can be safely published instead. The use of synthetic datasets is not new in research, and in fact has been suggested as a potential solution for some cases. First of all, in order to address the data disclosure limitation mentioned earlier, and particularly for national statistics providers whose main concern is to not reveal any individual identifying information (Rubin, 1993), researchers have suggested multiple processes for replicating data so that resulting datasets may be safely used instead (Raghunathan et al., 2003; Reiter, 2002). Secondly, further cases for the use of synthetic datasets have been made in order to provide highly controlled environments for the calibration of computational models (e.g. for training machine-learning algorithms) (Tomás et al., 2014). In addition to the above, this paper will present a third case for synthetic datasets, noting their use in the communication and reproducibility of computational geospatial models, especially in cases where the models have been built around 'un-publishable' datasets.

3. COMMUNICATING GEOSPATIAL MODELS

Computational geospatial models are often initially developed around a particular dataset in order to examine a particular phenomenon or observation, and may be (at least at first) built to work with a particular dataset schema as input. This is often particularly true for novel work, where not enough time has been invested yet to fully document the fine model mechanics and input data formats, but rather a model is often built to demonstrate proof-of-concept rather than satisfy widespread application. However, it is often at this stage that it is most important for modellers to publish their work, so that it may be examined and commented on by other members of the

community, and recommendations and improvements may be made while the work is still in active development and when feedback is most efficient.

3.1 The need for data publication

Regarding model publication, one part of the effective communication of a computational geospatial model to the wider research community requires the publication of the model in some form (be it through descriptive text, UML, or even better the code base itself), so that other researchers may run the model and experiment with it as well, as discussed Grimm and colleagues (Grimm, 2002; Grimm & Railsback, 2012). Furthermore, it is often important that a dataset needs to accompany the model publication as well, so that interested audiences may actually execute the computer model, examine its workings, and inspect it during runtime. Therefore, in a case where a model has been initially developed around an 'unpublishable' dataset, and furthermore whose comprehension and understanding requires input data, the 'unpublishable' dataset presents a significant bottleneck in the academic dissemination pipeline, as the model cannot be fully inspected without its input data, but the data may not be published along with the model due to external restrictions.

In these cases, it might be beneficial for the modellers to provide a synthetic dataset along with the presentation of their work, by means of constructing an artificial dataset that conforms to the data schema of the original dataset. In this way, the modellers avoid the disclosure of protected information, while at the same time providing a dataset which can be used as input to their model. Furthermore, a synthetic database may be more effective for initial model communication compared to the original dataset, as modellers have the capability to only include necessary attributes, thereby decluttering the dataset and better instructing audiences on what is required for the model to run.

3.2 Examples of synthetic geospatial data

As established earlier, publishing a dataset alongside a study requiring said dataset is beneficial for reader comprehension of the study in question. With the rise of data availability and its increasing use in research in recent years, the benefit of dataset dissemination is being acknowledged more and more by the academic community. This can be seen in publication outlets dedicated to the publication of high quality datasets, as is the journal *Data in Brief* for example, whose aim is the publication of citable datasets either for reference purposes tied to existing papers that make use of the dataset, or for potential future use by other researchers.

In the geospatial field in particular, a wide range of datasets are available, covering aspects from wildfire population risk (Mitsopoulos et al., 2020; Robinne, 2020), to drought occurrence and groundwater recharge sites (Dossou-Yovo et al., 2018; Rajasekhar et al., 2018), to socio-economic indicators (Mikhaylov et al., 2018), to urban design and built environment (Bartzokas-Tsiompras et al., 2021) illustrating the need for high quality publishable datasets. In regards to synthetic data, examples are found where synthetic datasets are published for use in machine learning algorithms (Sánchez & Vasile, 2020) as well as in the evaluation of algorithmic pipelines (Marelli et al., 2020); more relevant to the geospatial field, there are examples of detailed synthetic population datasets (Dennett et al., 2016; Joubert, 2018), of significant value to researchers wishing

to test or explore population dynamics at fine levels of details but are lacking access to real datasets.

3.3 Improving computational geospatial models through data publication

The previous section established the need and importance of including datasets, either real or synthetic, along with publications that rely on such datasets. This section discusses recent examples of computational geospatial models published in a relevant journal (the *European Journal of Geography*, in this case, for example see Mathioulakis & Photis, 2017; Photis & Sirigos, 2016; Fraile-Jurado et al., 2019; Zaleshina & Zaleshin, 2017; Bartzokas-Tsiompras & Photis, 2020; Paraskevopoulos et al., 2019). These publications are not accompanied by data, but rather present novel methodologies or applications of geospatial models through presentation of the underlying model. Four examples are briefly discussed regarding its aims, as well as limitations regarding the publication of data, and furthermore, for each example a short discussion follows outlining the potential benefits of publishing a sample dataset (synthetic or otherwise) along with the model for the particular paper.

Mathioulakis and Photis present an application of a geospatial model simulating future urban growth in the Greater Eastern Attica area in Greece (Mathioulakis & Photis, 2017). Their approach implements the standard SLEUTH model via a cellular automata (CA) model. CA models are a group of dynamic computational models operating on a spatial grid, with each cell in the grid evolving and changing its state over time based on inputs and a set of predetermined behavioural rules. CA are often set up via a programming language, and SLEUTH models in particular require a set of input data containing spatial characteristics (e.g. Slope, Landuse, Transportation, etc.), with input data provided in the form of digital image files at specified dimensions and encoding. The authors present the application process in great detail, however neither the code used to program the CA nor the input data are presented, therefore rendering the exact replication of the study near impossible. This example highlights an important benefit of sharing geospatial data alongside a model, where model execution requires input data at a very specific format (in this case, grayscale 8-bit GIF format images, all with same dimensions). While the authors include figures showcasing the model input image files, digital sample image files at the correct format could further aid reader comprehension, and in cases where the model code is also shared through an online repository (e.g. Github), can enable readers to execute the model themselves.

Photis and Sirigos present a location-allocation model to support decision-making in planning for ambulance allocation in the city of Volos, Greece (Photis & Sirigos, 2016). The model is presented through its set of equations in mathematical form, and its implementation and use is done through an interactive GIS interface. The model presented in this example was applied to a case study for the city of Volos, Greece, and its input data includes a road network as well as tabular data representing incidents requiring ambulance use for a specified period. Given the sensitive nature of the input data relating to patient information, it is expected that it cannot be made public, and thus this incident input data is not presented in the paper. As the authors describe in the paper the incident data contains a significant amount of additional information, including date, time, and location of the incident, and incident urgency, and furthermore the road network is broken down into 5 road categories with additional metadata (e.g.

uni- or bi-directional, service load, etc.), all of which constitutes information that can be presented very succinctly through a synthetic digital sample of each dataset, hosted online and shared in the paper.

Fraile-Jurado et al. present a comparison of different methods for assessing current and future coastal vulnerability due to sea-level rise in Europe (Fraile-Jurado et al., 2019). In their paper they present a comprehensive review of seven different models used to analyze and map the sea level. Of particular interest in this study is the listing of required input parameters for each reviewed model, as it gives an indication of the different variables that each model takes into account for its output. While it is not the authors' responsibility to educate readers how external software works or to detail software requirements, given that this paper constitutes a review of specialized software, a preview of sample datasets required for different software can enable readers to better comprehend the level of detail that each software operates at in comparison to others.

Zaleshina and Zaleshin present an overview of brain mapping techniques in neuroscience, and further illustrate the relevance of spatial data processing methodologies, as often found in geospatial fields, for neuroimaging applications (Zaleshina & Zaleshin, 2017). This paper draws some clear similarities in methodologies of two very distinct fields, and highlights how the two fields might benefit from each other. The authors present their case through detailed diagrams and figures highlighting matches and correspondences in both approaches, however an interesting point can be made here, regarding standard methods and practices in scientific fields: it is often the case that a method (be it a data format, algorithm, software, etc.) becomes a 'standard' in a field after many years of use and general consensus among the field practitioners that it is indeed a 'good' method. However this process carries with it all the minute nuances that each individual adds during the development and refinement of the method, and therefore even if two fields start from the same basic concept (e.g. the use of coordinate systems in both geosciences and neuroscience), it might be the case that after years of refinement each discipline has arrived at a significantly different form of the method. Within the context of this example, a sample dataset outlining the 'most common' attributes and data structures from each discipline could offer practitioners from each field a hands-on way of examining the ways that the other field approaches similar data.

4. GENERATING SYNTHETIC DATASETS

The previous sections established the benefits of publishing datasets along with the studies that rely on them, however as has been mentioned already this is often not feasible, due to publication constrains on datasets provided to researchers. In these latter cases of work based on 'un-publishable' data a solution has been identified in the use of synthetic data, i.e. made-up data points that have been generated to look like the real-world dataset, but do not actually describe a real-world case. This section will discuss in more detail the process and general aims of generating synthetic datasets, as well as the different levels of similarity a synthetic dataset may have in relation to the source dataset.

The process of generating a synthetic dataset has a seemingly straightforward aim, but the methods needed may become quite complex, quite quickly. The core aim is essentially to create an artificial dataset, which 'looks' and 'behaves' exactly like a

dataset that reflects the real-world. Using an original dataset, attributes are copied over to the new synthetic dataset and new values are assigned so that the resulting dataset is free of any relation to the real-world, and therefore does not contain any individual identifiers. While the aim is fairly straightforward, it is the level of similarity to the original dataset that may introduce increased complexity in the process. Using an original source dataset as a blueprint, three levels of similarity between the synthetic and the source dataset may be broadly identified (in ascending order of complexity): First: matching the source dataset's schema; second: matching the dataset's relational aspects; and third: introducing additional attribute relations.

In order to generate a synthetic dataset at any of the above levels of similarity to a source dataset, an understanding of the original dataset is needed, as the aim is to replicate aspects of it in some form. Furthermore, depending on the intended level of similarity to the original, proportionally deeper understanding of the original dataset is needed, as the aim becomes to replicate relationships between dataset attributes, e.g. correlations etc., which are rarely evident from a quick inspection of the original, and often require detailed analysis (and may in fact be the result of other, significant bodies of work). A graphical representation for the process of generating a sample synthetic dataset at varying levels of similarity is shown in **Figure 2**.

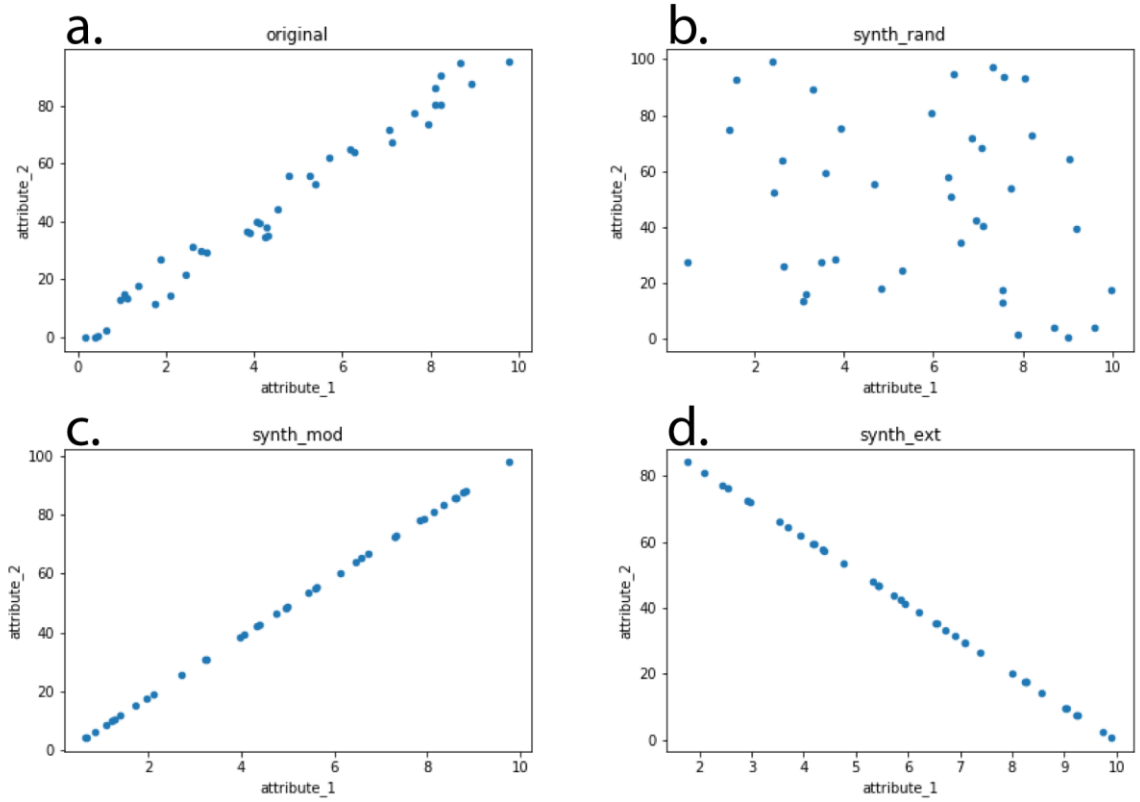

Figure 2. Synthetic generation of a sample dataset. Figure a represents the original (source) dataset, consisting of two positively correlated attributes (attribute_1, attribute_2), with values ranging between [0-10] for attribute_1 and [0-100] for attribute_2. Figure b represents a randomized synthesis, where the same attributes exist with values in the appropriate ranges, but no correlation is replicated. Figure c represents a more detailed synthesis, where in addition to characteristics of b, the correlation between attributes is also intrduced. Figure d represents an extension of the dataset, where non-existent characteristics are introduced (the negative correlation), for purposes of limit testing.

More specifically, for a given original dataset (**Figure 2a**) with two numeric variables (*attribute_1, attribute_2*) exhibiting a known correlation, synthetic datasets generated at each of the aforementioned levels of similarity would be the following: for the first level of similarity (**Figure 2b**), that of matching the *dataset schema*, the goal is to have a resulting dataset which contains all of the same attributes, each attribute has a value of the correct type, and furthermore values for each attribute are within the same range. Therefore, in a cursory computational check (i.e. through an algorithm that checks whether attributes exist) the resulting dataset will appear 'valid' and can indeed be fed into a relevant algorithm as input without any errors; however, *no condition is established on the relations between the attributes*, and in fact such a dataset has no analytical value, as all values have potentially been assigned at random. For synthetic datasets with analytical value, it is important *to match the relational aspects between the dataset's attributes*, that is respecting any correlations within the dataset (**Figure 2c**). This second level of similarity increases the complexity of the process significantly, as the dataset needs to be analysed beforehand, to identify correlations. At this point, if such a thorough analysis and replication is achieved, it could be assumed that the synthetic dataset is analytically equivalent to the original. The third level mentioned above deviates from the similarity process, and is included here as a theoretical end-point: if a synthetic dataset is generated to such a degree of similarity that it can be the equivalent of a real-world dataset and can be fed into a model, then it stands to reason that the process of generating relations in the dataset can be pushed further, to include correlations and relations not observed in the original, but also not prohibited by the relations in the data themselves (**Figure 2d**). Such a dataset may then be used to test extreme scenarios of the model, and examine the model's responsiveness to theoretically valid values, as a form of model sensitivity analysis.

5. AN EXAMPLE OF SYNTHETIC GEOSPATIAL DATASET GENERATION

Having discussed the reasoning and process through which synthetic datasets are generated, this section will present some examples on the use of synthetic data, used for the presentation of work done based on proprietary data. The work presented here constitutes the development of a dashboard platform, a set of non-specialist tools to be used for the analysis and visualisation of urban freight traffic. The proof-of-concept dashboard tool was built around a sample dataset of road freight traffic provided by a local authority in London, UK, as part of a partnership within the scope of the 'Freight Traffic Control 2050: Transforming the energy demands of last-mile urban freight through collaborative logistics' research project. As the original data was shared with a non-disclosure agreement, a synthetic dataset was constructed replicating the original, to demonstrate the dashboard tool's applicability. The dashboard development process is documented in (Cheliotis, 2019); the dashboard codebase is published under an open-source license and hosted at a public GitHub repository[2], along with the generated synthetic dataset. The synthetic dataset is constructed so that it fully matches the original's data schema, and furthermore some relational properties have been

[2] Dashboard GitHub repository found at: https://github.com/cheliotk/ftc2050-dashboard

maintained. Some samples of both cases will be presented, to better illustrate the differences.

5.1 Dataset generation process

Figure 3. Geospatial model study area extents, where synthetic trip data is generated for a particular local authority in London, UK. Greater London in light blue, local authority highlighted in red

The original dataset contains records of courier trips passing through a local authority in Central London, UK (**Figure 3**) for a period of one month. Daily activity patterns are quite consistent between days, with most of the activity taking place during business hours, and therefore it was decided that for the synthetic dataset, a single day would be generated.

The trips are categorized by trip origin and destination (OD) pair in relation to the area of interest (Internal/External), resulting in four categories: External-External (EE), External-Internal (EI), Internal-External (IE), and Internal-Internal (II). These categories served as the starting point for the generation of the synthetic data: In a Jupyter notebook[3] running python 3.4, point pairs were generated at random within the Greater London boundary using the geopandas[4] library, and categorized based on where they were located in relation to the local authority.

After generating the trip origin-destination coordinate pairs, trip path geometries on the street network were generated using the Open Source Routing Machine (OSRM)[5]. OSRM takes as input an OD pair and using a street network graph from OpenStreetMap calculates the shortest path between the two points, returning a GeoJSON object.

[3] https://jupyter.org/
[4] http://geopandas.org/
[5] http://project-osrm.org/

Following the trip path geometries, trips were filtered so that only geometries intersecting the area of interest (the local authority) were kept, matching the original dataset. Approximately 1000 trips were generated in total, and the number of trips for each category was chosen so that distributions were similar to those exhibited in the original dataset, as shown in **Table 1**. For each trip, the final trip path geometry was appended to the synthetic trip data to form a single trip record, which was then imported into a MongoDB[6] database to be retrieved by the web dashboard tool.

Table 1. Comparison of original and synthetic trip data by trip type in terms of origin-destination pair

	Original (1 day data)		Synthetic	
Trip type	**# of trips**	**%**	**# of trips**	**%**
External-External	3752	48.11%	497	50.92%
External-Internal	1619	20.76%	190	19.47%
Internal-External	1617	20.74%	190	19.47%
Internal-Internal	810	10.39%	99	10.14%
Total	7798	100.00%	976	100.00%

5.2 Levels of similarity to original

5.2.1 Case 1: No relational similarity

Trip origins and destinations were generated at random, therefore showing a distinct difference in clustering to actual destinations. In the original (**Figure 4** left), trip destinations are highly clustered around high-attractivity locations such as depots and train stations, while in the synthetic dataset (**Figure 4Error! Reference source not found.** right) trip destinations are randomly dispersed within the area, presenting a more homogeneous spatial distribution, with any discernible hotspots being the result of randomness, rather than spatial characteristics of the area.

Figure 4. Trip destinations heatmap. Original (left) and Synthetic (right)

[6] https://www.mongodb.com/

5.2.2 Case 2: Some relational similarity

The daily activity curve highlights the difference between business hours and off-hours traffic (**Figure 5**). The original shows a distinct difference between the two, with reduced activity before 5am, rising sharply from 6am through to 12am, and then gradually dropping again. While this has been replicated somewhat, the daily activity curve in the synthetic dataset has been modelled as a normal distribution, with a mean of 12am (noon) and standard deviation of 4.5 hours, therefore maintaining overall difference between night-day, but with a more gradual build-up and fall-off.

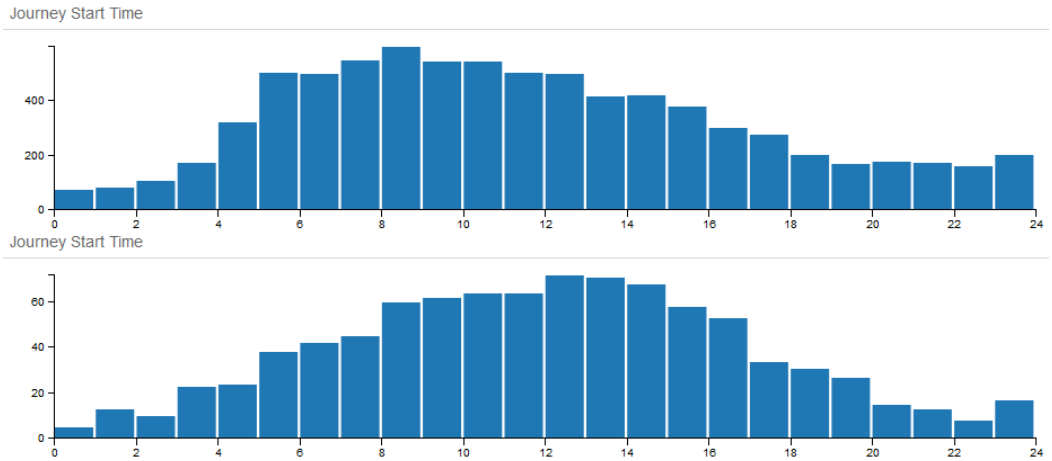

Figure 5. Daily activity curve (Number of rips starting per hour). Original (top) and Synthetic (bottom)

5.2.3 Case 3: Attribute relationship consistency

The original dataset contains two spatial identifiers for the trip's origin-destination location pair: one coordinate pair for the origin, one for the destination (both in WGS84), and an additional two-letter identifier for the whole trip signifying whether the trip started and/or ended within an area of interest (for a total of four distinct values). Generating each attribute at random would have resulted in potential inconsistencies between the string identifier and the actual trip origin and destination locations. The synthetic dataset strongly maintains the relationship between coordinates and the two-letter identifier, by essentially generating the string identifier based on the randomly generated point locations. An example of this consistency along with a comparison of trip data for a random record in the dataset is shown in **Figure 6**.

6. CONCLUSIONS

The inclusion of synthetic datasets for the communication of computational geospatial models provides some significant benefits: First of all, the generation of a new, artificial dataset that mimics an original dataset means that the information contained within does not reflect aspects of the real-world at all, is free of any disclosure limitations attached to the original, and can therefore be shared freely. Secondly, the process of generating an artificial dataset requires a more in-depth understanding of the original data, as well as the model, and allows the modeller to refine and declutter the synthetic

data shared to the wider community. And finally, through this process the modellers may generate additional datasets with characteristics not contained within the sampled original data, therefore allowing for more robust model development by exploring and pushing the model's capabilities further.

Figure 6. Attribute list comparison. Original (left) and Synthetic (right)

However, at the same time, the process of generating synthetic data has some significant drawbacks and limitations. A major issue is introduced by the additional workload required to produce the synthetic dataset, which may require significantly more time the more detailed the dataset is intended to be. And furthermore, given that a synthetic dataset at any adequate level of similarity to the original requires the modeller to first analyse and understand the original dataset they have in their hands, and therefore make assumptions on the relationships between existing attributes, it stands to reason that synthetic datasets may include bias: they are products of a modeller's assumptions and understanding of the underlying data, and therefore the modeller's bias may be well encrusted within the resulting synthetic dataset.

In conclusion, synthetic datasets may introduce biases through the assumptions included during the generation process, and may introduce additional workload for researchers especially during the early stages of model development leading to initial publications. But at the same time, if the model in question has been developed around datasets which came with disclosure limitations, often the generation and inclusion of a synthetic dataset might be the only way of including integral operational data for a more efficient communication and reproducibility of the model.

NOTES

The dataset and work mentioned in this project (FTC2050 dashboard) is made available to explore on Github: https://github.com/cheliotk/ftc2050-dashboard

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the EPSRC for funding this work through its financial support of Freight Traffic Control 2050 (www.ftc2050.com), EPSRC Grant Reference: EP/N02222X/1. Responsibility for the contents of the paper rests with the authors.

REFERENCES

Angus, S. D., & Hassani-Mahmooei, B. (2015). "Anarchy" Reigns: A Quantitative Analysis of Agent-Based Modelling Publication Practices in JASSS, 2001-2012. *Journal of Artificial Societies and Social Simulation*, *18*(4), 16.

Aspinall, P., Mavros, P., Coyne, R., & Roe, J. (2013). The urban brain: Analysing outdoor physical activity with mobile EEG. *British Journal of Sports Medicine*, bjsports-2012-091877. https://doi.org/10.1136/bjsports-2012-091877

Bartha, G., & Kocsis, S. (2011). Standardization of geographic data: The european inspire directive. *European Journal of Geography*, *2*(2), 79–89.

Bartzokas-Tsiompras, A., & Photis, Y.N. (2020). Does neighborhood walkability affect ethnic diversity in Berlin? Insights from a spatial modeling approach, *European Journal of Geography*, 11 (1), pp. 163-187, https://doi.org/10.48088/ejg.a.bar.11.1.163.187

Bartzokas-Tsiompras, A., Photis, Y. N., Tsagkis, P., & Panagiotopoulos, G. (2021). Microscale walkability indicators for fifty-nine European central urban areas: An open-access tabular dataset and a geospatial web-based platform. *Data in Brief*, 36, 107048. https://doi.org/10.1016/j.dib.2021.107048

Calabrese, F., Colonna, M., Lovisolo, P., Parata, D., & Ratti, C. (2011). Real-Time Urban Monitoring Using Cell Phones: A Case Study in Rome. *IEEE Transactions on Intelligent Transportation Systems*, *12*(1), 141–151. https://doi.org/10.1109/TITS.2010.2074196

Cheliotis, K. (2019). *Developing a Dashboard of Last-Mile Freight Traffic* (Internal Report No. 1; p. 10). University College London. http://www.ftc2050.com/reports/ftc2050-dashboardReport-final.pdf

Dennett, A., Norman, P., Shelton, N., & Stuchbury, R. (2016). A synthetic Longitudinal Study dataset for England and Wales. *Data in Brief*, *9*, 85–89. https://doi.org/10.1016/j.dib.2016.08.036

Diao, M., Zhu, Y., Ferreira, J., & Ratti, C. (2015). Inferring individual daily activities from mobile phone traces: A Boston example. *Environment and Planning B: Planning and Design*, 0265813515600896. https://doi.org/10.1177/0265813515600896

Dittus, M., Quattrone, G., & Capra, L. (2017). Mass Participation During Emergency Response: Event-centric Crowdsourcing in Humanitarian Mapping. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1290–1303. https://doi.org/10.1145/2998181.2998216

Dossou-Yovo, E. R., Kouyaté, A. M., Sawadogo, T., Ouédraogo, I., Bakare, O. S., & Zwart, S. J. (2018). A geospatial database of drought occurrence in inland valleys in Mali, Burkina Faso and Nigeria. *Data in Brief*, *19*, 2008–2014. https://doi.org/10.1016/j.dib.2018.06.105

Fraile-Jurado, P., Iglesias-Campos, A., Simon-Colina, A., & Hodgson, N. (2019). Methods for assessing current and future coastal vulnerability to sea level rise. A review for a case-study in Europe. *European Journal of Geography*, *10*(3), 97–119.

Grimm, V. (2002). Visual Debugging: A Way of Analyzing, Understanding and Communicating Bottom-up Simulation Models in Ecology. *Natural Resource Modeling*, *15*(1), 23–38. https://doi.org/10.1111/j.1939-7445.2002.tb00078.x

Grimm, V., Berger, U., Bastiansen, F., Eliassen, S., Ginot, V., Giske, J., Goss-Custard, J., Grand, T., Heinz, S. K., Huse, G., Huth, A., Jepsen, J. U., Jørgensen, C., Mooij, W. M., Müller, B., Pe'er, G., Piou, C., Railsback, S. F., Robbins, A. M., … DeAngelis, D. L. (2006). A standard protocol for describing individual-based and agent-based models. *Ecological Modelling*, *198*(1–2), 115–126. https://doi.org/10.1016/j.ecolmodel.2006.04.023

Grimm, V., Berger, U., DeAngelis, D. L., Polhill, J. G., Giske, J., & Railsback, S. F. (2010). The ODD protocol: A review and first update. *Ecological Modelling*, *221*(23), 2760–2768. https://doi.org/10.1016/j.ecolmodel.2010.08.019

Grimm, V., & Railsback, S. F. (2012). Designing, Formulating, and Communicating Agent-Based Models. In A. J. Heppenstall, A. T. Crooks, L. M. See, & M. Batty (Eds.), *Agent-Based Models of Geographical Systems* (pp. 361–377). Springer Netherlands. https://doi.org/10.1007/978-90-481-8927-4_17

Haklay, M., & Weber, P. (2008). OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Computing*, *7*(4), 12–18. https://doi.org/10.1109/MPRV.2008.80

Joubert, J. W. (2018). Synthetic populations of South African urban areas. *Data in Brief*, *19*, 1012–1020. https://doi.org/10.1016/j.dib.2018.05.126

Kitchin, R. (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. SAGE Publications Ltd. http://methods.sagepub.com/book/the-data-revolution

Liu, L., Andris, C., & Ratti, C. (2010). Uncovering cabdrivers' behavior patterns from their digital traces. *Computers, Environment and Urban Systems*, *34*(6), 541–548. https://doi.org/10.1016/j.compenvurbsys.2010.07.004

Marelli, D., Bianco, S., & Ciocca, G. (2020). IVL-SYNTHSFM-v2: A synthetic dataset with exact ground truth for the evaluation of 3D reconstruction pipelines. *Data in Brief*, *29*, 105041. https://doi.org/10.1016/j.dib.2019.105041

Mathioulakis, S., & Photis, Y. N. (2017). Using the sleuth model to simulate future urban growth in the Greater Eastern Attica area, Greece. *European Journal of Geography*, *8*(2), 107–120.

Mikhaylov, A. S., Mikhaylova, A. A., & Kuznetsova, T. Yu. (2018). Geospatial dataset for analyzing socio-economic regional divergence of European regions. *Data in Brief*, *19*, 2374–2383. https://doi.org/10.1016/j.dib.2018.07.027

Mitsopoulos, I., Mallinis, G., Dimitrakopoulos, A., Xanthopoulos, G., Eftychidis, G., & Goldammer, J. G. (2020). Vulnerability of peri-urban and residential areas to landscape fires in Greece: Evidence by wildland-urban interface data. *Data in Brief*, *31*, 106025. https://doi.org/10.1016/j.dib.2020.106025

Pampel, H., & Dallmeier-Tiessen, S. (2014). Open Research Data: From Vision to Practice. In S. Bartling & S. Friesike (Eds.), *Opening Science: The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing* (pp. 213–224). Springer International Publishing. https://doi.org/10.1007/978-3-319-00026-8_14

Paraskevopoulos, Y., Bardosa, A., Photis, Y.N. (2019). Exploring the impact of network configuration and transport accessibility on population dynamics. The case of Naxos island, Greece, *European Journal of Geography*, 10 (4), pp. 177-194

Photis, Y. N., & Sirigos, S. A. (2016). Scenario-based location of ambulances for spatiotemporal clusters of events and stochastic vehicle availability. A decision support systems approach. *European Journal of Geogrpahy*, *Volume 6*(No 4).

Raghunathan, T. E., Reiter, J. P., & Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, *19*(1), 1.

Rajasekhar, M., Sudarsana Raju, G., Siddi Raju, R., & Imran Basha, U. (2018). Data on artificial recharge sites identified by geospatial tools in semi-arid region of Anantapur District, Andhra Pradesh, India. *Data in Brief*, *19*, 462–474. https://doi.org/10.1016/j.dib.2018.04.050

Reades, J., Calabrese, F., Sevtsuk, A., & Ratti, C. (2007). Cellular Census: Explorations in Urban Data Collection. *IEEE Pervasive Computing*, *6*(3), 30–38. https://doi.org/10.1109/MPRV.2007.53

Reades, Jonathan, Calabrese, F., & Ratti, C. (2009). Eigenplaces: Analysing Cities Using the Space–Time Structure of the Mobile Phone Network. *Environment and Planning B: Planning and Design*, *36*(5), 824–836. https://doi.org/10.1068/b34133t

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of O¢ Cial Statistics*, 531–544.

Robinne, F.-N. (2020). A geospatial dataset providing first-order indicators of wildfire risks to water supply in Canada and Alaska. *Data in Brief*, *29*, 105171. https://doi.org/10.1016/j.dib.2020.105171

Rubin, D. B. (1993). Statistical disclosure limitation. *Journal of Official Statistics*, *9*(2), 461–468.

Sánchez, L., & Vasile, M. (2020). Synthetic database of space objects encounter events subject to epistemic uncertainty. *Data in Brief*, *32*, 106298. https://doi.org/10.1016/j.dib.2020.106298

Tomás, J. T., Spolaôr, N., Cherman, E. A., & Monard, M. C. (2014). A Framework to Generate Synthetic Multi-label Datasets. *Electronic Notes in Theoretical Computer Science*, *302*, 155–176. https://doi.org/10.1016/j.entcs.2014.01.025

Zaleshina, M., & Zaleshin, A. (2017). The brain as a multi-layered map. Scales and reference points for pattern recognition in neuroimaging. *European Journal of Geography*, *8*(1), 6–31.