A reproducible approach to generating synthetic spatial data for teaching and learning purposes

Paddy Gorry*[1] and Peter Mooney†[1,2]

[1]Hamilton Institute, Maynooth University, Ireland
[2]Department of Computer Science, Maynooth University, Ireland

GISRUK 2023

Summary

Whilst there is an ever increasing amount of openly available spatial-data for teaching and learning purposes teachers and students often wish to generate their own synthetic spatial datasets. Such datasets can be used to test algorithms, test software code, assist in visualisation development, and be used as unseen datasets for student assessment and learning. We describe the development of a Python-based software tool called RADIAN ((**RA**n**D**om spat**I**al d**A**ta ge**N**erator)) for generating simple synthetic spatial datasets.

KEYWORDS: Synthetic data, teaching and learning, spatial data, Python, reproducibility.

1 Introduction and motivation

In some teaching and learning scenarios it is often necessary to access randomly-generated or synthetic spatial datasets for student assessment and learning and software or visualisation testing. Providing students with interesting (spatial) datasets during teaching and learning activities in classes can be challenging. However, their provision can improve student engagement (Burns and Chopra, 2016; Dhanorkar et al., 2021) and help address privacy and ethical risks (Berg et al., 2016) around certain datasets. Generating sample datasets for the classroom from open or public datasets (Tosadori et al., 2016; Calderini and Harding, 2019) is possible but is often a manual process and can be very time consuming (Bart et al., 2017). Being able to generate synthetic datasets quickly is an attractive requirement particularly in fast-moving and dynamic classroom environments. Studies have shown that students engage better with teachers and instructors who show originality and ingenuity in developing their own teaching materials (Anderson et al., 2021; Rao and Dave, 2019). In this paper we describe the design, development and implementation of a reproducible Python-based software tool called RADIAN (**RA**n**D**om spat**I**al d**A**ta ge**N**erator). This

*patrick.gorry.2015@mumail.ie
†peter.mooney@mu.ie

software generates synthetic spatial datasets containing randomly-generated data quickly and easily without the need for external datasets. RADIAN has been specifically developed for use within teaching and learning contexts involving spatial data and GIS. RADIAN allows specification of the characteristics of a synthetic spatial dataset including geographic extent, number of geographic features, number of attributes per feature, and the specific data types for each attribute.

2 Related Work

Collection, analysis, and visualization of spatial data are essential components of spatial literacy and spatial thinking (Jarvis, 2011). Iterative and interactive approaches to teaching spatial data analysis facilitate deeper engagement with concepts (Kedron et al., 2022) while making learning of spatial data handling techniques interesting and relevant encourages successful student engagement (Welle Donker et al., 2022; Gunderman, 2020). However, as discussed by Bearman et al. (2016) many GIS modules often spend more time developing the technical skills associated with using ArcGIS, QGIS and other GIS software, rather than developing theoretical understanding of spatial problems. Several efforts have been made to create software specifically for spatial data science teaching and learning usage. Calderini and Harding (2019) describe *GRD* in R which provides students with an custom-tailored data sets for statistical exploration. Mannino and Abouzied (2019) describe a tool that helps users generate "real-looking synthetic data" by specifying the properties of the dataset. QGIS provides functionality where randomly generated points can be generated. For a given spatial extent parameters control minimum spacing between points, total number of points generated, and so on. Only an ID attributed is generated for each spatial object.

3 RADIAN implementation

RADIAN is implemented in Python and uses all open source Python libraries for spatial data handling including GeoPandas[1], Shapely[2], Sklearn[3] and geovoronoi[4]. RADIAN is available as open source software via GitHub [5] allowing researchers or students to use the tool for their own synthetic data generation purposes. As input RADIAN accepts a GeoJSON file representing the outer boundary of the region of interest. All of the parameters for RADIAN can be configured easily via a JSON file. Upon execution, RADIAN automatically, in a unsupervised fashion, generates point-based data with the characteristics within the specified boundary. RADIAN uses several approaches to generate these points including Voronoi-based buffering by localising generation of points within these Voronoi-polygons (Gold et al., 1996). RADIAN can also choose to generate points within Voronoi or buffer regions based on the overall area of the local region and distance from the local or global centroid. The idea here is to avoid simply dividing the geographical space evenly across the input region and placing random points within these divisions. Readers who wish to find a full detailed explanation of the methodology behind RADIAN's data generation are

[1] https://geopandas.org/en/stable/

[2] https://shapely.readthedocs.io/en/stable/manual.html

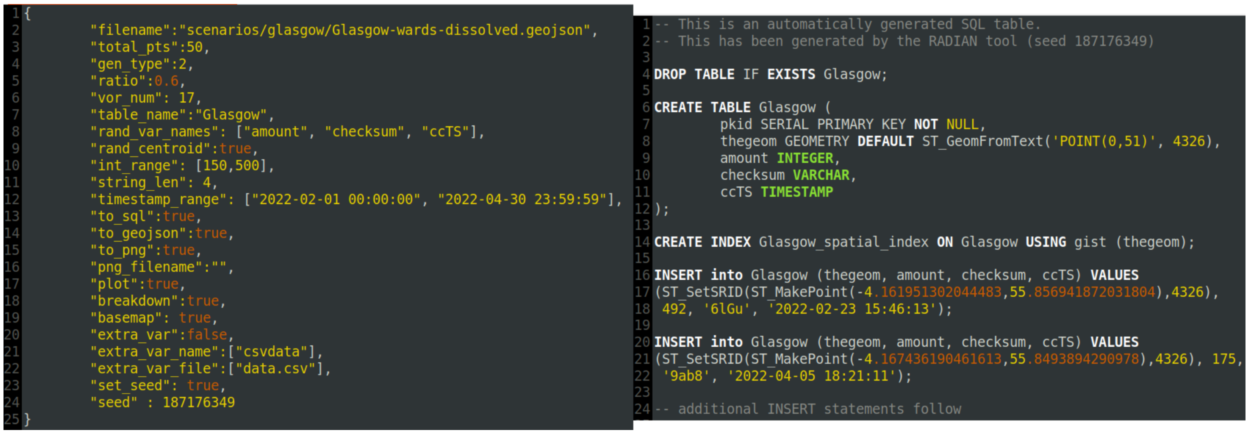
[3] https://scikit-learn.org/stable/

[4] https://github.com/WZBSocialScienceCenter/geovoronoi

[5] https://github.com/paddeaux/msc_rng

referred to the open-access MSc thesis publication available at Gorry and Mooney (2022).

RADIAN outputs the spatial dataset as a GeoJSON file and as a PostgreSQL dump file (suitable for import directly into a PostgreSQL PostGIS-enabled database). Both of these files can then be disseminated to students as learning materials or used by the teacher in assessments. The data generated by RADIAN allows teachers to create datasets that "look" realistic, are appropriately contextualized for the target student group or activity, and are illustrative of deeper learning concepts (Kim et al., 2018). RADIAN allows specification of attributes having timestamps within ranges, randomly generated strings of characters, primitive data types and the selection of values from predefined code-lists of values (stored in CSV files). A `seed` value can be specified allowing for greater reproducability in the generated datasets while also allowing students to replicate exact datasets using the same `seed` and parameter values. Figure 1 shows (left) the JSON configuration file for RADIAN using a GeoJSON boundary of Glasgow city wards as the outer boundary of the region of interest. This configuration corresponds directly to Figure 2 for 50 randomly generated point objects representing fictional smartphone-based online transactions in this region. A PostgreSQL PostGIS compliant SQL is generated (right of figure 1) where the randomly generated attribute values are shown in the `INSERT` statements. Figure 3 shows a similiar synthetic dataset generated by RADIAN and QGIS. A very similar configuration file is used for RADIAN. QGIS does not generate any additional attributes for objects and output files must be manually exported from the QGIS interface. The main difference between the QGIS and RADIAN generated points is RADIAN's preference on generating points around clusters without striving to distribute points evenly within the selected region.



```
{
        "filename":"scenarios/glasgow/Glasgow-wards-dissolved.geojson",
        "total_pts":50,
        "gen_type":2,
        "ratio":0.6,
        "vor_num": 17,
        "table_name":"Glasgow",
        "rand_var_names": ["amount", "checksum", "ccTS"],
        "rand_centroid":true,
        "int_range": [150,500],
        "string_len": 4,
        "timestamp_range": ["2022-02-01 00:00:00", "2022-04-30 23:59:59"],
        "to_sql":true,
        "to_geojson":true,
        "to_png":true,
        "png_filename":"",
        "plot":true,
        "breakdown":true,
        "basemap": true,
        "extra_var":false,
        "extra_var_name":["csvdata"],
        "extra_var_file":["data.csv"],
        "set_seed": true,
        "seed" : 187176349
}
```

```
-- This is an automatically generated SQL table.
-- This has been generated by the RADIAN tool (seed 187176349)

DROP TABLE IF EXISTS Glasgow;

CREATE TABLE Glasgow (
        pkid SERIAL PRIMARY KEY NOT NULL,
        thegeom GEOMETRY DEFAULT ST_GeomFromText('POINT(0,51)', 4326),
        amount INTEGER,
        checksum VARCHAR,
        ccTS TIMESTAMP
);

CREATE INDEX Glasgow_spatial_index ON Glasgow USING gist (thegeom);

INSERT into Glasgow (thegeom, amount, checksum, ccTS) VALUES
(ST_SetSRID(ST_MakePoint(-4.161951302044483,55.856941872031804),4326),
 492, '6lGu', '2022-02-23 15:46:13');

INSERT into Glasgow (thegeom, amount, checksum, ccTS) VALUES
(ST_SetSRID(ST_MakePoint(-4.167436190461613,55.8493894290978),4326), 175,
 '9ab8', '2022-04-05 18:21:11');

-- additional INSERT statements follow
```

Figure 1: On the left the parameters configuration JSON file for RADIAN is shown related to the output in Figure 2. On the right the output PostgreSQL PostGIS compliant SQL file is shown for two point objects. The object attribute values are shown in the `INSERT` statements

4 Conclusions and future work

We successfully deployed RADIAN within an MSc Postgraduate course on Spatial Databases during September to December 2022 in Maynooth University. Feedback has been very positive. The key

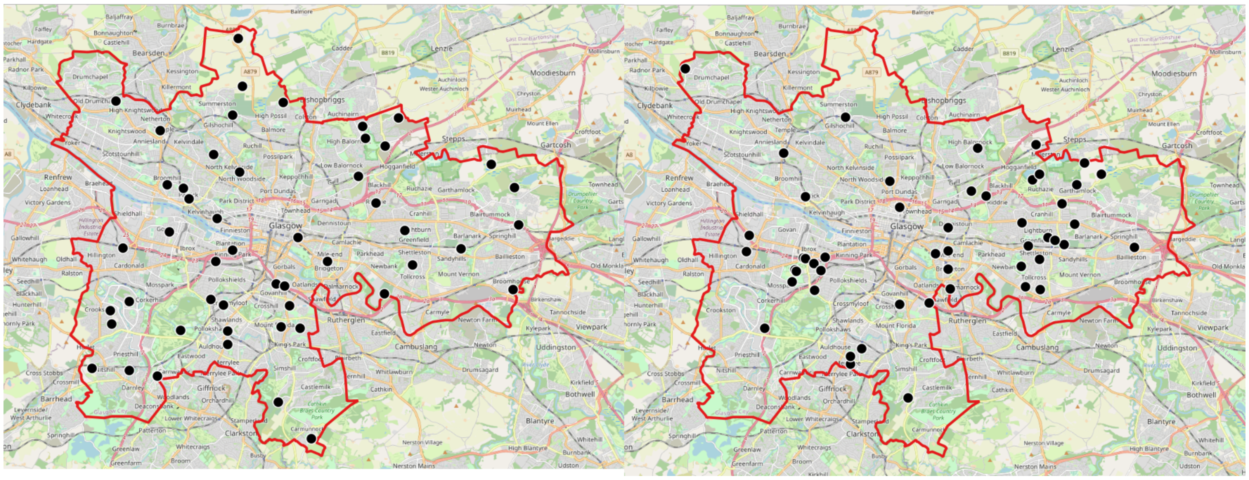


Figure 2: On the left QGIS generates 50 random points within a polygon boundary of several district wards around Glasgow, Scotland. The image on the right is produced by RADIAN generating 50 random points within the same boundary

strengths of RADIAN in this deployment included the ability to generate datasets for classroom assessment quickly and generating *interesting* synthetic datasets for teaching key concepts (for example a synthetic geocoded Instagram-photos dataset for a city in teaching heatmaps and kernel density estimation). In our full paper presentation at GISRUK 2023 we will outline a number of interesting and exciting developments around our work on RADIAN including results of a user survey around the usability and real-world similarity of RADIAN generated datasets and the first steps towards integration of Artificial Intelligent approaches (such as GANs (Cunningham et al., 2022)) in the future development of RADIAN.

Acknowledgements

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6049. The opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Science Foundation Ireland.

Biography

Paddy Gorry is a 1st year PhD student in the Hamilton Institute at Maynooth University. RADIAN was developed during his MSc in Data Analytics in 2022. Paddy is pursuing a PhD in Computer Science in the area of using Artificial Intelligence approaches for random (spatial) data generation. Peter Mooney is an Assistant Professor of Computer Science at Maynooth University. He is supervising Paddy's PhD research. Peter is also a member of the GISRUK National Steering Committee.

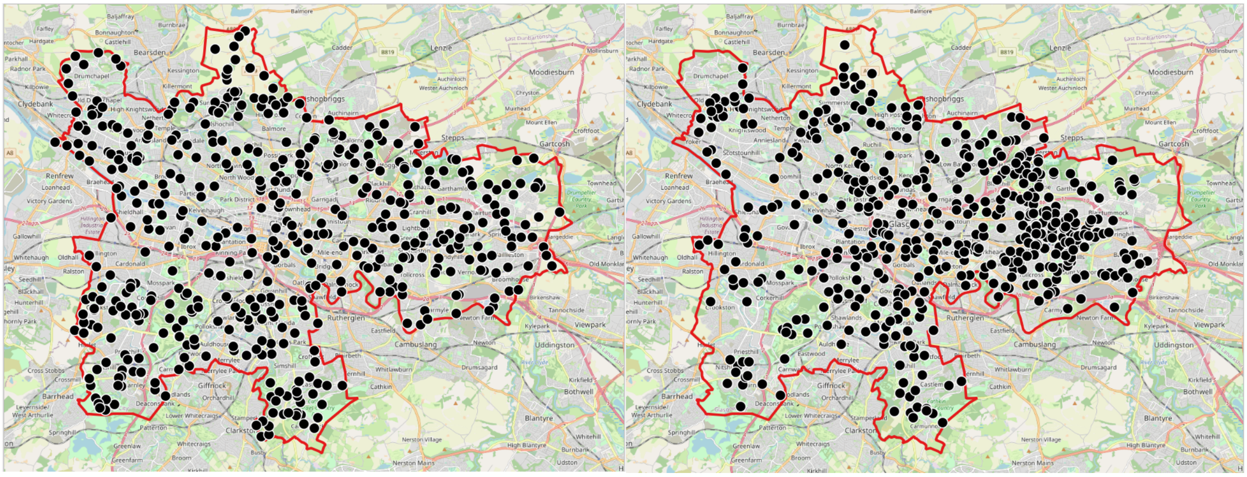


Figure 3: On the left QGIS generates 500 random points within a polygon boundary of several district wards around Glasgow, Scotland. The image on the right is produced by RADIAN generating 500 random points within the same boundary

References

Anderson, R. C., Bousselot, T., Katz-Buoincontro, J., and Todd, J. (2021). Generating buoyancy in a sea of uncertainty: Teachers creativity and well-being during the covid-19 pandemic. *Frontiers in Psychology*, 11:614774.

Bart, A. C., Whitcomb, R., Kafura, D., Shaffer, C. A., and Tilevich, E. (2017). Computing with corgis: Diverse, real-world datasets for introductory computing. *ACM Inroads*, 8(2):66–72.

Bearman, N., Jones, N., André, I., Cachinho, H. A., and DeMers, M. (2016). The future role of gis education in creating critical spatial thinkers. *Journal of Geography in Higher Education*, 40(3):394–408.

Berg, A. M., Mol, S. T., Kismihok, G., and Sclater, N. (2016). The role of a reference synthetic data generator within the field of learning analytics. *Journal of Learning Analytics*, 3(1):107–128.

Burns, C. and Chopra, S. (2016). A meta-analysis of the effect of industry engagement on student learning in undergraduate programs. *The Journal of Technology, Management, and Applied Engineering*, 33(1).

Calderini, M. and Harding, B. (2019). Grd for r: An intuitive tool for generating random data in r. *The Quantitative Methods for Psychology*, 15(1):1–11.

Cunningham, T., Klemmer, K., Wen, H., and Ferhatosmanoglu, H. (2022). Geopointgan: Synthetic spatial data with local label differential privacy. *arXiv preprint arXiv:2205.08886*.

Dhanorkar, S., Rosson, M. B., and Hellar, D. B. (2021). Exploring techniques for promoting engagement with lecture content: A synthetic review. *EDULEARN21 Proceedings*, pages 9622–9632.

Gold, C. M., Remmele, P. R., and Roos, T. (1996). Voronoi methods in gis. *Advanced School on the Algorithmic Foundations of Geographic Information Systems*, pages 21–35.

Gorry, P. and Mooney, P. (2022). Random generation of realistic spatial data for use in classroom assessments. M.Sc thesis available as PDF at `https://mural.maynoothuniversity.ie/16871/`.

Gunderman, H. C. (2020). Developing lesson plans for teaching spatial data management in academic libraries through a lens of popular culture. *Journal of Map and Geography Libraries*, 16(3):239–253.

Jarvis, C. H. (2011). Spatial literacy and the postgraduate gis curriculum. *Procedia - Social and Behavioral Sciences*, 21:294–299. International Conference: Spatial Thinking and Geographic Information Sciences 2011.

Kedron, P., Quick, M., Hilgendorf, Z., and Sachdeva, M. (2022). Using the specification curve to teach spatial data analysis and explore geographic uncertainties. *Journal of Geography in Higher Education*, 46(2):304–314.

Kim, A. Y., Ismay, C., and Chunn, J. (2018). The fivethirtyeight r package:'tame data'principles for introductory statistics and data science courses. *Technology Innovations in Statistics Education*, 11(1).

Mannino, M. and Abouzied, A. (2019). Is this real? generating synthetic data that looks real. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, UIST '19, page 549–561, New York, NY, USA. Association for Computing Machinery.

Rao, A. R. and Dave, R. (2019). Developing hands-on laboratory exercises for teaching stem students the internet-of-things, cloud computing and blockchain applications. In *2019 IEEE Integrated STEM Education Conference (ISEC)*, pages 191–198. IEEE.

Tosadori, G., Bestvina, I., Spoto, F., Laudanna, C., and Scardoni, G. (2016). Creating, generating and comparing random network models with networkrandomizer. *F1000Research*, 5.

Welle Donker, F., van Loenen, B., Keßler, C., Küppers, N., Panek, M., Mansourian, A., Zhou, P., Vancauwenberghe, G., Tomić, H., and Kević, K. (2022). Showcase of active learning and teaching practices in spatial data infrastructure (sdi) education. *AGILE: GIScience Series*, 3:1–11.