

Classification of Missing Geospatial Data from Structure and Mechanism Perspective

Merve Polat Kayali^{*1}, Ana Basiri^{†1}

School of Geographical & Earth Sciences, University of Glasgow

Summary

Data-centric science, data-empowered society, and policymaking based on data can suffer from flawed conclusions if data are representative, biased, or unavailable. This paper focuses on missingness for which the common mitigation and handling strategies is a deletion or single imputation. However, understanding the reasons causing the missingness can help to understand phenomena better. Distinguishing the different types of missingness help us to develop and implement new imputation approaches, sampling strategies and output uncertainty quantification. In this paper, using missing data mechanism and structure a new taxonomy has been created to classify the causalities of missing geospatial data.

KEYWORDS: missing data, taxonomy of missingness, geospatial data

1. Introduction

Missing data is the lack of observational data in an acquired dataset (Kang, 2013). In general, missing data problems can be encountered in all fields of science. It poses a significant challenge to the results of statistical studies which needed to use in different fields. Missing data in the datasets due to the destruction of the documents in which the data is included, incorrect data recordings, wrong calculations by the analysts (Gong et al., 2021), and many unpredictable natural events create some problems in the statistical analysis of the data. The reason why the missing data in the observed datasets is very important is the possibility that these data may have a crucial effect on the process or outcome of the research. Due to missing data, important data may be hidden that completely changes the meaning of statistical analyses. We need to distinguish different types of missingness to be able to develop and implement different mitigation plans, imputation approaches.

2. Classification of Missing Data

Hand (2020) uses the term ‘Dark Data’ to refer to different sorts of missing data. This term originated from the physics comparison of dark matter. In his book, a taxonomy is used to explain and classify the many different reasons for missing data. This taxonomy has 15 types of Dark Data to represent missing data or inadequate that can occur (Hand, 2020). This system of classification is useful because it was developed to provide a specific checklist of threats and general difficulties while looking at any dataset.

Classification of missing data types may be limited when analysing the causes of this data. This classification is based on the data itself or observed data. Rubin (Rubin, 1976) describes a theory that is the basis for the classification of missing data. Based on this study, missing data has been classified into three types: missing at random (MAR), missing completely at random (MCAR), not at missing

^{*} m.polat-kayali.1@research.gla.ac.uk

[†] Ana.Basiri@glasgow.ac.uk

random(NMAR)(Little and Rubin, 2019). Suppose there is missing data for a variable in a dataset. If the probability of missing data is independent and unrelated to itself or other observed data, this missing data is missing at completely random (MCAR). If the missing data is not related to its value but is missing due to other values observed in the dataset, it is missing at random (MRA)(Allison, 2001). If the reason for missing data is due to the values of the variable itself, then missing data is missing at not random (NMAR).

Spatial data heterogeneity and the first law of geography suggest that closer phenomena are more likely to be similar (Tobler, 1970). Therefore it is plausible that missingness that is not at random follows a spatial pattern if the nature of the variable is spatial. This can potentially help to understand the underlying reasons why missingness happened in the first place. In the light of studies and classifications, it has been examined for what reasons the missing data may appear on a geospatial basis. This section has analysed the classification of missing data. The next part of this paper will describe a taxonomy that has been created based on whether the missing data is known or not known.

3. Taxonomy of Spatial Missing Data

This paper proposes a taxonomy of missing geospatial data. As it is illustrated in Figure1 the taxonomy is based on whether the reasons for the missing data are known or unknown. So, the first category includes the types of missingness where we know what is missing, the second category includes we don't know what we don't know. The first category should be handled in two ways as knowing or not knowing the reason for these missing data. In the second category, we don't even know that we have missing data. In these cases, data loss may not be noticed. Especially in rural areas, errors may occur in geocoding due to some location access or low population. If the errors that occur are not understood by checking the dataset, the missing data in the geocoding areas cannot be noticed.

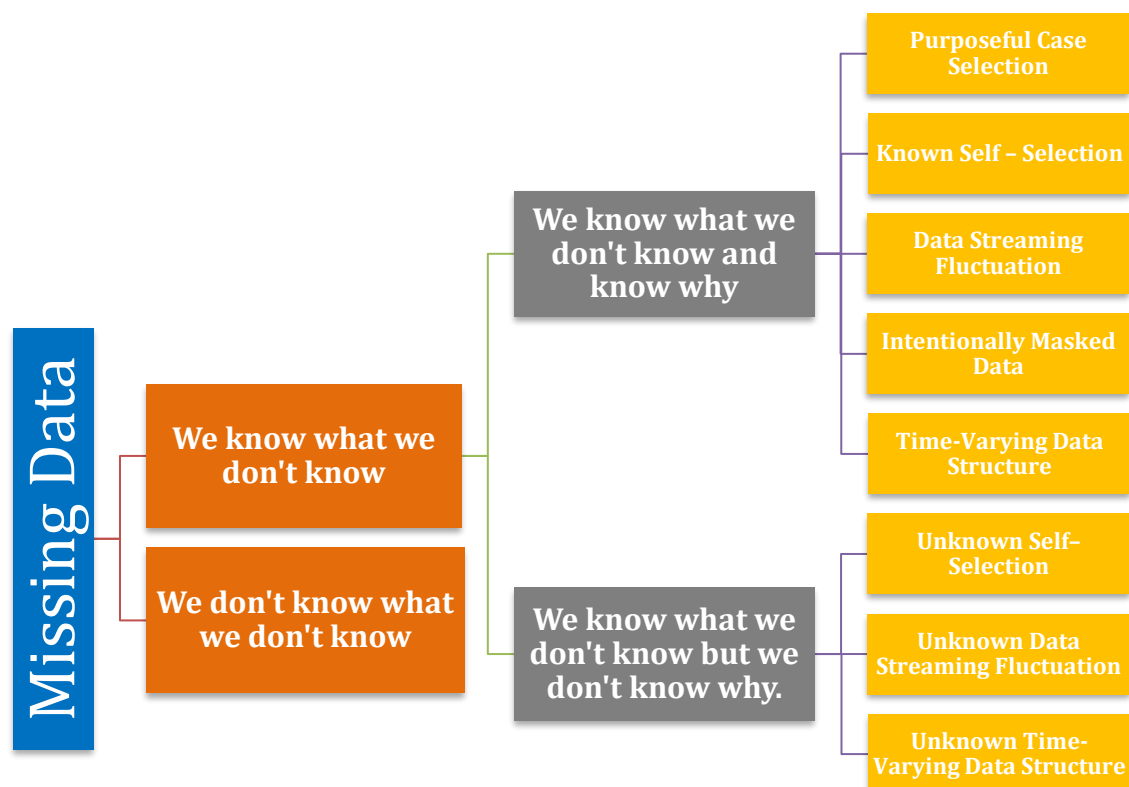


Figure 1. Overview of the Taxonomy

3.1. We know what we don't know

3.1.1. We know what we don't know and know why

These are the cases that we don't know the values of the missing data, but we know the situations that cause the missing data. There may be some pollution in the data set we have. For example, data in multispectral satellite images contaminated or missing by the presence of clouds must be reconstructed (Lorenzi et al., 2013). We know that the data collected in this example is missing or polluted by clouds.

3.1.1.1. Purposeful Case Selection

It is the case of selecting only data that match certain criteria among the data that can be included in a sample. Although these choices may seem appropriate at the beginning, they will not reflect the real values at the end of the study.

3.1.1.2. Known Self-Selection

In this situation, people choose whether to include their data in the dataset, so studies with this data may not be completely correct. Especially when dealing with data on immigrants, there may be some missing data when recording the number of immigrants at certain observation intervals. Data in the survey may be incomplete as a result of unanswered questions by immigrants (Rogers et al., 2010).

3.1.1.3. Data Streaming Fluctuation

All measurement processes may have missing data or errors caused by changes in measurement conditions or random fluctuations. Data may be collected incorrectly or incompletely due to instantaneous physical problems in data collection devices. For example, missing data caused by beam blockage, failure of devices, and near-ground blind areas are often observed when using weather radars. Beam blockage occurs when land or other objects such as houses and trees block the radar beam and a wedge-shaped blind zone emerges behind the item (Geiss and Hardin, 2021).

3.1.1.4. Intentionally Masked Data

In this situation, only some data are selected and others are not specifically included in the dataset. There may be data that has been intentionally obscured to ensure the privacy and security of the location. For instance, location-based services take users' location data and offer them products and services near them. However, it poses a security and privacy threat as intruders can easily learn about users' instant location (Kachore et al., 2015). In these cases, location obfuscation for location data privacy occurs.

3.1.1.5. Time-Varying Data Structure

Data can be changed and hidden in many ways over time. Some changes may occur after the observation periods of the data. Definitions of data may change over time for different reasons, depending on the situation it is in. For instance, a comprehensive geographic information system (GIS) dataset of vector-based surface water features such as streams, lakes, and rivers may be incomplete or time-dependent. The flow rate of the streams changes depending on the weather conditions in certain periods of the year (Kim, 2018).

3.1.2. We know what we don't know but we don't know why

Missing data can be caused by different reasons such as self-selection, measurement error, or time-varying.

3.1.2.1. Unknown Self-Selection

Missing data from observed data is not known to be missing due to self-selection. To illustrate, we have missing location data in a survey for women who live in high crime rates locations, because they don't want to give their address information detail, but we don't know the reason for the missing address data we have.

3.1.2.2. Unknown Data Streaming Fluctuation

This situation arises because it is not known that the reason for missing data after data collection is due to measurement errors. For example, during the measurement of high-frequency radars, beam extinction which occurs because of strong storms can result in significant missing data zones (Geiss and Hardin, 2021).

3.1.2.3. Unknown Time-Varying Data Structure

Some data may change over time. If the values obtained while observing a dataset change depending on time or definition, there may be deficiencies in these data.

4. Conclusion and Future Works

Missing data is a growing challenge in time of big data and so categorising the structure and mechanism causing it is the first step in having a better understanding of the problem and potential solutions. Future work can include more informative imputation, missing data visualisation, output uncertainty quantification based on input missingness. These will help a wide range of disciplines including geospatial data science, mapping, and visualisation, as well as applied areas of research including epidemiology and social sciences to appreciate the impact of using missingness and the potential uncertainty associated with as useful data for their inference.

5. Acknowledgements

Great thanks for the financial support Republic of Turkey Ministry of National Education Scholarship Program. The work presented in this paper has been funded by the UK Research and Innovation (UKRI) Future Leaders Fellowship MR/S01795X/2.

References

- ALLISON, P. D. 2001. *Missing data*, Sage publications.
- GEISS, A. & HARDIN, J. C. 2021. Inpainting Radar Missing Data Regions with Deep Learning. *Atmospheric Measurement Techniques Discussions*, 1-28.
- GONG, Y., LI, Z., ZHANG, J., LIU, W., CHEN, B. & DONG, X. A spatial missing value imputation method for multi-view urban statistical data. *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 2021. 1310-1316.
- HAND, D. J. 2020. *Dark data: Why what you don't know matters*, Princeton University Press.
- KACHORE, V. A., LAKSHMI, J. & NANDY, S. K. 2015. Location Obfuscation for Location Data Privacy. *2015 IEEE World Congress on Services*.
- KANG, H. 2013. The prevention and handling of the missing data. *Korean journal of anesthesiology*, 64, 402.

- KIM, D. H. 2018. High-spatial-resolution streamflow estimation at ungauged river sites or gauged sites with missing data using the National Hydrography Dataset (NHD) and U.S. Geological Survey (USGS) streamflow data. *Journal of Hydrology*, 565, 819-834.
- LITTLE, R. J. & RUBIN, D. B. 2019. *Statistical analysis with missing data*, John Wiley & Sons.
- LORENZI, L., MELGANI, F. & MERCIER, G. 2013. Missing-Area Reconstruction in Multispectral Images Under a Compressive Sensing Perspective. *IEEE Transactions on Geoscience and Remote Sensing*, 51, 3998-4008.
- ROGERS, A., LITTLE, J., RAYMER, J. & SPRINGERLINK 2010. *The indirect estimation of migration: methods for dealing with irregular, inadequate, and missing data*, New York;Dordrecht;, Springer.
- RUBIN, D. B. 1976. Inference and missing data. *Biometrika*, 63, 581-592.
- TOBLER, W. R. 1970. A computer movie simulating urban growth in the Detroit region. *Economic geography*, 46, 234-240.

Biographies

Merve Polat Kayalı is a Ph.D. student in Geospatial Data Science at the University of Glasgow. Her research interests are missing data, geospatial data science, augmented reality.

Ana Basiri is a Professor of Geospatial Data Science and a UKRI Future Leaders Fellow at the University of Glasgow.