IDFC Institute

DGN Policy Brief 09

March 2021

Data Governance Network
Anchored by IDFC Institute

Assessing geomasking techniques to protect the spatial privacy of individuals during COVID-19 and beyond

By Harsh Vardhan Pachisia, Isalyne Gennaro and Kadambari Shah

State governments have ramped up their use of Geographic Information Systems (GIS) during the COVID-19 pandemic to track the spread of the virus. However, the use of individual addresses and subsequent published maps showing these locations has rightly brought up issues of spatial privacy. This policy brief summarises various geomasking techniques that can be applied by Indian policymakers to reduce safety and security concerns. With such measures, GIS can continue to be used not only for tracking COVID-19 spread but also for monitoring and improving the delivery of healthcare beyond the pandemic.

Context

Geographic Information System (GIS) is a technology that combines multiple layers of information for a spatial understanding of various patterns. GIS has played a significant role in fighting the COVID-19 crisis across the world and in India.

As early as March 2020, India's central government agencies such as National Informatics Centre and National Disaster Management Authority worked with GIS firms to plan, coordinate and implement various action plans to curb the spread of the novel coronavirus. Healthcare centres, quarantine facilities and testing camps were added to GIS platforms, allowing officials to track the whereabouts of infected/suspected COVID-19 positive individuals, identify potential clusters of infection and initiate countermeasures (John, 2020).

State and local governments also employed GIS in their pandemic control and relief efforts. For instance, Kerala and Gujarat's state governments, and Srinagar and Mumbai's local administrations launched different GIS-based mobile apps/platforms to identify high-risk areas and conduct contact tracing.

GIS is an important tool to improve policymaking, beyond the pandemic. It can be applied to understand health concerns, integrate healthcare schemes and identify beneficiaries, deploy additional medical resources, track progress and cure rates, and so on, for more efficient delivery of health services. At the same time, however, the release of this information poses a threat to confidentiality and individual privacy. For instance, through a process called 'reverse geocoding', publishing the address or coordinates (latitude/longitude) of an individual infected with COVID-19 makes it possible to discover the name of the individual, age, date of birth, etc. by showcasing the actual address and then linking it to other available directories. Given that leading GIS platforms such as ArcGIS have inbuilt reverse geocoding tools, it becomes critical to protect the geo privacy of individuals when publishing such datasets.

Geomasking protection strategies

Spatial data of locations/addresses of individuals when tracking diseases should not be released unless the data points have been altered using either aggregation or geographic masking. Aggregating such data to the ward or city level is often useful when monitoring diseases at a macro-level. However, it strips away the ability to conduct point-to-point analysis (i.e. between two locations) and determine where granular disease clusters are being formed. Geographic masking overcomes this limitation. Franch-Pardo et. al. (2020) showed multiple researchers using GIS-based methods that require granular locations to track and mitigate the spread of COVID-19 in real-time. The researchers employed geomasking techniques that decrease the likelihood of re-identification without aggregating the data to a coarser unit of analysis.

Techniques

Zandbergen, P. A. (2014) summarises some geomasking techniques (Figure 1). Each of the five techniques contain the general steps:

Figure 1: Geomasking techniques highlighted in Zandbergen (2014)

1. The aim of each technique is to add a degree of randomisation and noise to the dataset to reduce the risk of re-identification through reverse geocoding.
2. The original locations are randomised/ displaced within a fixed radius, on the circumference of a circle. Some techniques such as 'Donut Masking' and 'Bimodal Gaussian Displacement' conduct more complex variations such as setting a minimum amount of displacement (the radius of the circle) or having the amount of displacement following a normal distribution.
3. All original locations are replaced with masked/displaced locations in the dataset that is eventually published.

To further enhance security, Zhang et.al (2017) propose the 'location swapping' method that allows only existing residential addresses to become candidates for masked locations rather than any location within the circle.

Though there are no universally accepted guidelines, it is widely agreed that the amount of displacement should be inversely proportional to the local population density. Intuitively, one would understand that if there are more people in a given area, a small possible area will encapsulate enough of them to make identifying one person difficult, such as a densely populated slum; conversely, if the area is sparsely populated, a larger possible area will be required to encompass enough people to make identifying one person difficult, such as a rural area where the population is less or more dispersed.

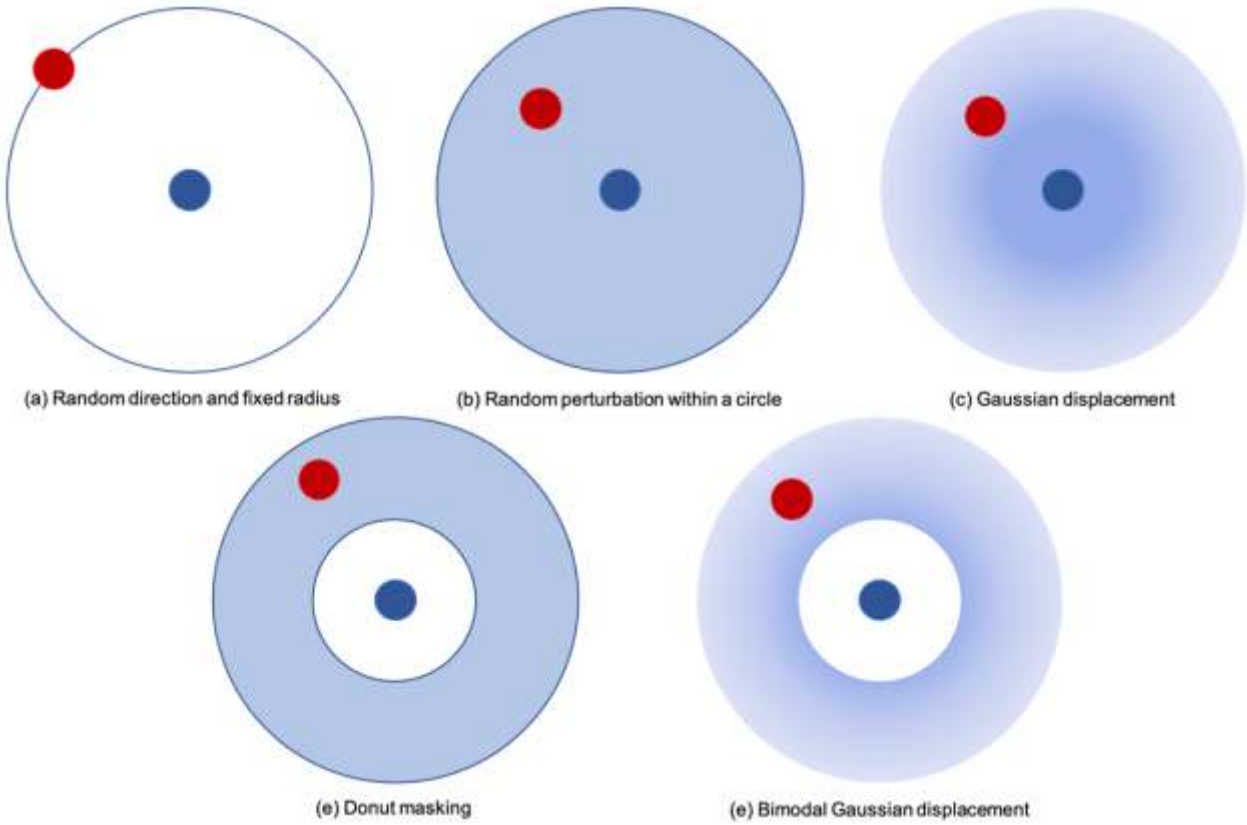
For example, when the government wants to protect the location of people receiving health coverage under nation-wide healthcare programmes such as the Ayushman Bharat Yojana, a high degree of displacement should be applied in rural areas where the location of beneficiaries can easily be re-identified. This is especially important if a large part of the area chosen covers a forest in which presumably no one resides.

Limitations of such techniques

1. The masked location is still an address of a human being whose privacy is being invaded and might be wrongly scrutinised when they are not an actual part of the study/scheme.
2. These techniques are not easy to use. The policymaker still has to determine the radius of displacement before using such a technique and obtain granular population data.

To overcome these limitations, Swanlund et. al. (2020) propose 'street masking', a

Figure 1: Geomasking techniques highlighted in Zandbergen (2014)



Source: Authors visualisation based on Zandbergen, P. A. (2014)

technique that masks the original location by moving it to a node on the street network (instead of another address) through a series of calculations.

Effectiveness of such techniques

How can policymakers measure the effectiveness of each technique to mask the data? This can be achieved by working out the probability of discovery even after geomasking is undertaken.

One measure to calculate this is proposed by Samarati and Sweeney (1998), and is called *k-anonymity.* A dataset is *k-anonymous* if it has a minimum of *k-1* data points (people) with the same characteristics when anonymisation has been completed. If *k* is not sufficiently high, more anonymisation is necessary. This concept has been expanded to spatial data, called *spatial k-anonymity,* where one would use addresses (geographic identifiers) to link the spatial datasets. *Spatial k-anonymity* can be defined as the number of addresses from which the original location cannot be distinguished. In this case, reverse geocoding becomes the primary mechanism for re-identification. An empirical estimate of spatial k-anonymity can be achieved by employing 'the nth nearest neighbour number' method. A series of calculations gives the number of potential addresses which are closer to the masked address than the original location (Allshouse et.al., 2010).

Generating a high degree of spatial k-anonymity is critical in measuring the effectiveness of geomasking the data. However, it should be kept in mind that there are issues with spatial k-anonymity and it is vulnerable to certain forms of attack. Ultimately, no means of security is perfect but using geomasking techniques and maximising the degree of spatial k-anonymity is crucial when working with spatial datasets to track the spread of a disease such as COVID-19.

Recommendations

There is no consensus on which geomasking method is best suited for which kind of data. For example, when tracking the spread of COVID-19 and monitoring the enforcement of home quarantine by collecting data on infected individuals, any such methods could have been used. However, the government authorities could select a technique based on the degree of geomasking they wish to attain in order to protect an individual's privacy. It is crucial that policy makers understand how each technique differs and its limitations:

- Donut masking and bimodal gaussian displacement are usually preferred. To mask the data point corresponding to the actual address of an individual, they add higher degrees of randomness and enforce a minimum amount of displacement in the dataset.
- In comparison, the street masking technique proposed by Swanlund et. al. (2020) protects the location of individuals even more. They place the masked location at a node of the street network, rather than at an individual address. However, to be effective, the technique requires good quality street network data. This may not be the case in rural areas of India and other developing countries. It could also be a challenge in densely-built urban areas with intricate street networks such as slum settlements. Moreover, this technique protects geographic information of individuals but groups or communities living in a particular neighbourhood can still be re-identified. This can be an issue in settlements where individuals of a same socio-economic or ethnic community live together, and could be exposed to discrimination or stigmatisation.

Before using individual-level geospatial data, researchers and policymakers must define the extent to which granular location data must be anonymised, and find the balance between the usefulness of publishing such data versus protecting the privacy of individuals in the dataset. This decision can be taken based on the established law on sensitivity of personal data and protection of spatial privacy of individuals, and a principles framework that incorporates considerations such as ethical soundness and technical feasibility. Ultimately, what matters is that policy makers employ 'confidentiality protection methods' that can reduce the risk for re-identification while preserving the properties of the spatial dataset itself, and ensure that GIS tools are used for decision-making safely.

References

1) Allshouse, W. B., Fitch, M. K., Hampton, K. H., Gesink, D. C., Doherty, I. A., Leone, P. A., ... & Miller, W. C. (2010). Geomasking sensitive health data and privacy protection: an evaluation using an E911 database. Geocarto international, 25(6), 443-452.

2) Franch-Pardo, I., Napoletano, B. M., Rosete-Verges, F., & Billa, L. (2020). Spatial analysis and GIS in the study of COVID-19. A review. Science of The Total Environment, 140033.

3) Jaiswal, P. B. Privacy of COVID-19 suspects violated; names, addresses made public. (2020). Accessed: https://www.theweek.in/news/india/2020/03/22/privacy-of-covid-19-suspects-violated-names-addresses-made-public.html

4) John, C. This Technology Is Helping India Fight Against Coronavirus. (2020). The Quint. Accessed: https://www.thequint.com/tech-and-auto/tech-news/what-is-gis-and-how-it-is-helping-battle-coronavirus-in-india

5) Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). l-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD), 1(1), 3-es.

6) Samarati, P., & Sweeney, L. (1998, June). Generalizing data to provide anonymity when disclosing information. In PODS (Vol. 98, No. 10.1145, pp. 275487-275508).

7) Song, F., Ma, T., Tian, Y., & Al-Rodhaan, M. (2019). A New Method of Privacy Protection: Random k-Anonymous. IEEE Access, 7, 75434-75445.

8) Swanlund, D., Schuurman, N., Zandbergen, P., & Brussoni, M. (2020). Street masking: a network-based geographic mask for easily protecting geoprivacy. International journal of health geographics, 19(1), 1-11.

9) Zandbergen, P. A. (2014). Ensuring confidentiality of geocoded health data: assessing geographic masking strategies for individual-level data. Advances in medicine, 2014.

10) Zhang, S., Freundschuh, S. M., Lenzer, K., & Zandbergen, P. A. (2017). The location swapping method for geomasking. Cartography and Geographic Information Science, 44(1), 22-34.

Data Governance Network

The Data Governance Network is developing a multi-disciplinary community of researchers tackling India's next policy frontiers: data-enabled policymaking and the digital economy. At DGN, we work to cultivate and communicate research stemming from diverse viewpoints on market regulation, information privacy and digital rights. Our hope is to generate balanced and networked perspectives on data governance - thereby helping governments make smart policy choices which advance the empowerment and protection of individuals in today's data-rich environment.

About Us

IDFC Institute has been set up as a research-focused think/do tank to investigate the political, economic and spatial dimensions of India's ongoing transition from a low income, state-led country to a prosperous market-based economy. We provide in-depth, actionable research and recommendations that are grounded in a contextual understanding of the political economy of execution. Our work rests on three pillars – 'State and the Citizen', 'Strengthening Institutions', and 'Urbanisation'. The State and the Citizen pillar covers the design and delivery of public goods, ranging from healthcare and infrastructure to a robust data protection regime. The Strengthening Institutions pillar focuses on improving the functioning and responsiveness of institutions. Finally, the Urbanisation pillar focuses on the historic transformation of India from a primarily rural to largely urban country. All our research, papers, databases, and recommendations are in the public domain and freely accessible through www.idfcinstitute.org.

About the Author

Harsh Vardhan Pachisia is an Associate at IDFC Institute. His research focuses on projects related to urban development, data science and governance. Prior to joining IDFC Institute, he worked as an Economics Teaching Fellow at the NYU Stern School of Business. Harsh has also been a Research Assistant to several NYU professors analysing topics such as 'Shadow Banking around the World' and 'Social Media Entrepreneurship'.

Isalyne Gennaro is an Associate at IDFC Institute. Her research interests lie in urban development, state capacity and the social sector in developing countries. She worked as an urban planner at HCP Design Planning and Management in Ahmedabad where she worked on policymaking for various public projects such as the Eastern Waterfront Redevelopment in Mumbai. She also did research on land utilisation in Indian cities published in Journal of Urban Design.

Kadambari Shah is a Senior Associate at IDFC Institute. Her research focuses on ease of doing business in India, state capacity, and urban governance. At St. Xavier's College, Mumbai, she was on the editorial board of several scholarly journals and has presented research papers on development issues ranging from the middle-income trap to the role of race and gender in electoral politics.

Acknowledgments

The authors would like to thank Vikram Sinha for valuable feedback, comments and editorial inputs. We are grateful to Girish Gupta whose prior work at IDFC Institute formed the basis of this policy brief. We are indebted to the Rockefeller Foundation, without whom this work would not have been possible.

Disclaimer and Terms of Use

IDFC Institute

301, 3rd Floor, Construction House 'A', 24th Road, Off Linking Road,
Khar West, Mumbai 400052

/idfcinstitute @idfcinstitute /IDFCInstitute