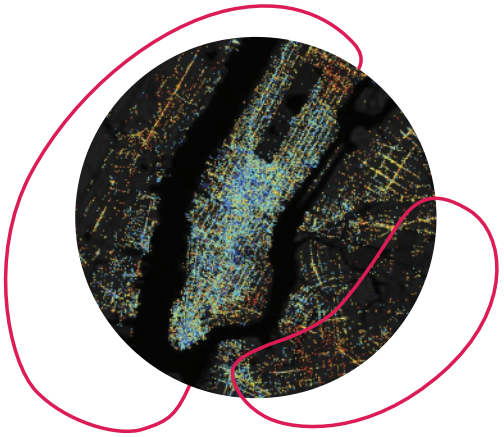
spectus™
Data. In sight.



WHITEPAPER

Differential Privacy

As part of our commitment to a privacy-by-design development framework, Spectus is constantly innovating geospatial privacy solutions that are also firmly aligned with our core values of Client Centricity and Transparency. In addition to the multiple governance and technical solutions we have already developed for the Spectus Data Clean Room, we are now adding an additional state-of-the-art solution to our privacy toolkit: true Differential Privacy.

Through our participation in Microsoft's SmartNoise Early Adopter Program, Spectus is now able to deliver more granular aggregated insights, thereby improving the utility of aggregated data without ever sacrificing the privacy of users who entrust Spectus with their data.

What it is

Differential privacy (DP) is a mathematical definition of privacy. A dataset is said to be differentially private if we can statistically bound the amount of individual-level information an attacker can deduce by looking at the dataset. This is achieved via a variety of mechanisms that add noise to the process generating the dataset itself.

The crucial advantage of differential privacy is that adding noise to these processes ensures that sensitive data is thoroughly protected and any potential malice is mitigated—and we can mathematically quantify how much.

Microsoft SmartNoise and OpenDP

Overview

Our datasets are generated using the SmartNoise differential privacy toolkit, which is developed by Microsoft and Harvard University, and is part of the OpenDP open-source software project.

As pioneers in Differential Privacy, Microsoft and Harvard's Institute for Quantitative Social Science (IQSS) and School of Engineering and Applied Sciences (SEAS) created OpenDP in order to make differential privacy more accessible to data scientists and practitioners. This collaboration has led to the creation of SmartNoise, an open-source differential privacy platform.

SmartNoise Early Adopter Acceleration Program

A Microsoft SmartNoise blog from December 2020 explains the reasoning driving the establishment of the SmartNoise Early Adopter Acceleration Program. John Kahan, VP and Chief Data Analytics Officer at Microsoft, wrote, “This collaboration program with the SmartNoise team aims to accelerate the adoption of differential privacy in solutions today that will open data and offer insights to benefit society.”

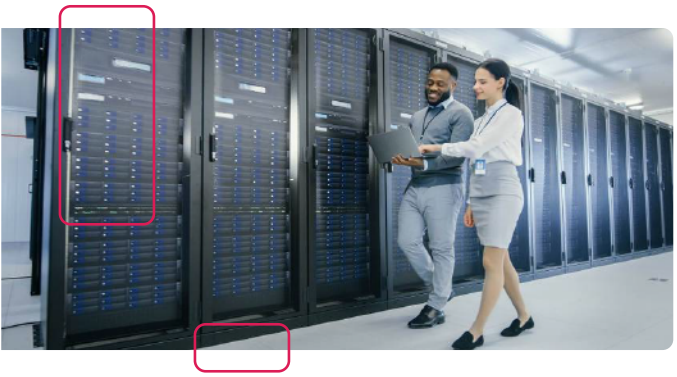
OpenDP

OpenDP is a community effort to build a trustworthy suite of open-source tools enabling privacy-protective analysis of sensitive personal data. The ultimate outcome is a library of algorithms that data scientists can use to generate differentially private statistical releases.

Why we participated

Our team is dedicated to learning and expanding our safety measures for ourselves, our partners, and our users. Becoming a member of this program allows us to make actionable, aggregated data more widely available to the research and humanitarian communities, while also providing Spectus with yet another technical solution to add to our privacy toolkit.

Spectus’ Data Governance



In addition to the multiple governance layers defining our ethical data collection and responsible data sharing framework, the Spectus Data Clean Room offers a suite of technical solutions to protect user privacy, while preserving the utility of the data. Not only does each solution approach privacy preservation in a different way, but they also have unique advantages based on the needs of the analysis at hand. These include:

Privacy Enhanced Mobility Data (PEM)

In order to achieve greater data utility without sacrificing user privacy, Spectus developed Privacy Enhanced Mobility Data (PEM), which introduces noise to the inferred home location of users to prevent re-identification. By improving already anonymized data with PEM, Spectus seeks to increase user privacy preservation while retaining the ability to infer broad user demographics, and to observe mobility behaviors of truly anonymous users across “whitelisted” points of interest—public and commercial venues that do not reveal sensitive information about individual users. Privacy Enhanced Mobility data for custom geospatial analyses and methods development is available within the Spectus Data Clean Room.

Highly Aggregated Data

By aggregating data to a high spatial resolution, such as a county, Spectus is able to provide turn-key insights for a number of mobility-related metrics. This includes trends like visitation rates to points of interest over time, origin-destination travel patterns, and evacuation rates during natural disasters.

During the aggregation process, certain steps are taken to ensure there are sufficient users within a given area to guarantee true anonymity. A key advantage of highly aggregated data is that it can be shared with a broad audience and can be easily analyzed without requiring a high degree of data science skills. However, given its highly aggregated nature, certain nuances may be lost when looking at higher spatial levels of granularity.

Differentially Private Aggregated Data

By applying differential privacy to the aggregation process, it is possible to aggregate at more granular levels of spatial resolution—such as municipality, census tract, and census block group—without compromising user privacy. Increased granularity allows decision-makers to better understand nuanced mobility behavior.

Including error risk

Introducing noise in the aggregation process presents the potential for error in data sets. In essence, adding noise to our data means sacrificing some accuracy in order to protect sensitive information. By including transparent error values within the differentially private datasets, we are able to effectively communicate caveats to downstream analysts of this data, so that they can take such errors into account within their decision making processes.

How we applied Differential Privacy for Social Good

As a pioneer of the “Data for Good” movement, Spectus drives positive social impact through the ethical and responsible use of location-based data. Spectus has incorporated DP methodologies to certain aggregated datasets, such as evacuation rates during natural disasters, in order to improve the utility of these analyses without sacrificing user privacy.

Evacuation rates

During natural disasters, such as hurricanes and wildfires, it is critical to understand the rate at which residents are evacuating impacted areas and where they are going. By providing these insights, our aim is to support emergency managers and public planners to better prepare for and respond to natural disasters.

Metrics Calculated

Spectus’ Differentially Private Evacuation Rate dashboard reports the percentage of residents who have evacuated their homes by county, the most common destinations, and the average distance traveled by evacuees. Furthermore, Spectus analyzes evacuation rates by income group in order to provide insights into the impact of natural disasters on various income segments.

> Some metrics need more protection than others.

For each county in the US of interest during the impacted timeframe, three metrics were calculated. Using Harris County,Texas as an example, the provided metrics are:

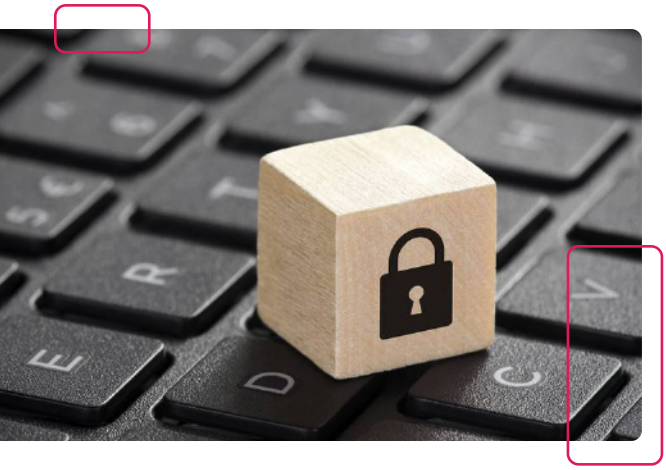
- **Evacuation Rate:** The percentage of people evacuated from Harris County, aka Harris evacuees;
- **Popular Destinations:** a list of destination counties and the percentage of Harris evacuees in each one of the destinations;
- **Average Distance From Home:** the average distance traveled from home by the Harris evacuees.

Noise added in aggregating operations

We add noise when generating and aggregating our metrics to ensure the data respects differential privacy constraints.

There are four different aggregating operations related to these metrics:

- **COUNT** of total users in Harris (CTU)
- **COUNT** of evacuees in Harris (CE)
- **COUNT** of Harris evacuees per destination (CED)
- **SUM** of distances from home values [in miles] of all the Harris evacuees (SD)



Privacy budget

The privacy budget (ε) is the amount of noise—and therefore the level of privacy guarantees—added by DP, which is represented by a real value. The higher the privacy budget, the more accurate the output dataset will be, but the more likely it is to release potentially private information.

In this case, the Laplace mechanism is used for all four operations. Finally, the privacy budget, ε, is set as follows:

- $\varepsilon = 4$ for CTU
- $\varepsilon = 2$ for CE
- $\varepsilon = 1$ for CED
- $\varepsilon = 3$ for SD

The rationale behind this is that some metrics need more protection than others. For example, the total users count (CTU) is less sensitive than the actual destinations of the users (CED) and thus we can use a higher privacy budget, which means higher accuracy.

After adding noise with the SmartNoise library according to these parameters the set of four DP values — CTU_{DP}, CE_{DP}, CED_{DP}, SD_{DP} — can be used to compute our private metrics.

- **evacuation_rate** = CE_{DP}/CTU_{DP}
- **popular_destinations** = CED_{DP}/CE_{DP}
- **avg_distance_from_home** = SD_{DP}/CTU_{DP}

Errors

Two different approaches have been used to estimate the errors introduced by the DP methodology. Specifically, a theoretical method was used for the avg_distance_from_home metric; while a simulation-based method has been used for the count-based metrics, namely the evacuation_rate and the popular_destinations. The estimation of the error is defined as the 95th percentile of the error distribution.

Theoretical Error - avg_distance_from_home

The noise distribution added to the output of an operation depends only on the sensitivity of the operation and the set privacy budget, $\mathcal{E}$. In particular, it is proven that adding noise according to the Laplace distribution centered in 0 and with:

$$scale = \frac{sensitivity}{\mathcal{E}}$$

The sensitivity of the AVG operation is (MAX-min)/N, where MAX and min are relative to the column we are computing the average on, and N is its size. Because of the filters on the diagonal bounding box column on the source table of our process, MAX and min are fixed (10^6 and 0 respectively).

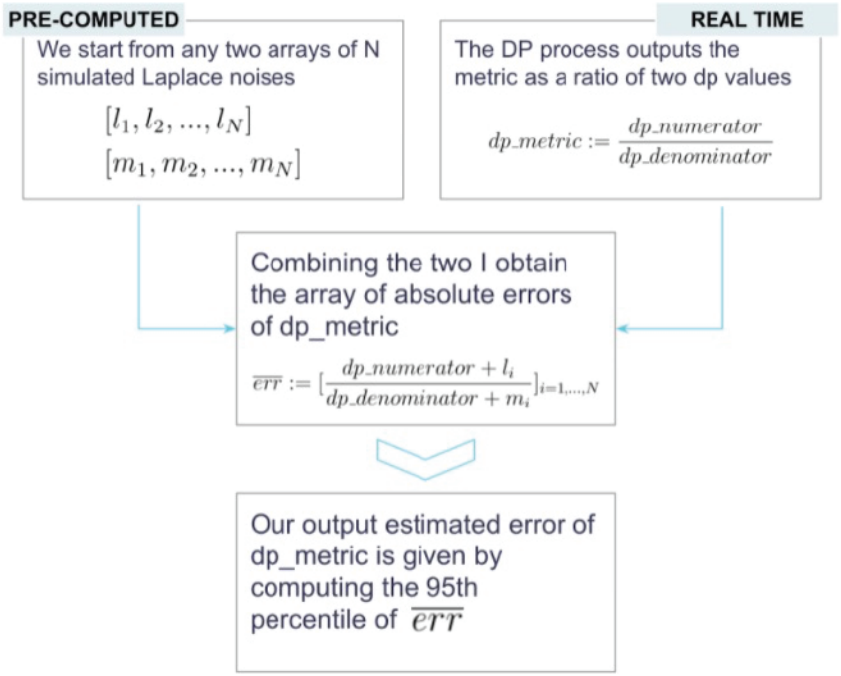
If you consider that in this case $\mathcal{E}$ is a fixed constant, all this means that the noise distribution for the avg_distance_from_home metric depends exclusively on N, the size of the aggregation bucket. Adding in the linearity of the percentile of the Laplace distribution with respect to its scale, the 95th percentile of the noise is computed as follows:

$$\begin{aligned} 95th\_perc\_error &= percentile(abs(Laplace(\frac{MAX - min}{N \cdot \epsilon})), 95) \\ &= \frac{1}{N} \cdot percentile(abs(Laplace(\frac{MAX - min}{\epsilon})), 95) \\ &= \frac{1}{N} \cdot C \end{aligned}$$

Where C is a constant that can be pre-computed. This can then be applied to each row of the output dataset, using CTUDP as N.

Simulating the error - evacuation_rate and popular_destinations

Since these metrics are both ratios, the error distribution is given by the ratio of DP mechanisms. This cannot be easily estimated mathematically and so a simulation approach was used. This is achieved by "caching" the noise added by the Laplace mechanisms (two for each statistic, one for the denominator and one for the numerator) and to reuse it at runtime to compute error estimates. This has to be done for each row, i.e. each pair (county, income segment), using the same pre-computed noise arrays.



Conclusion

By incorporating differential privacy to certain aggregated datasets, we are able to uncover trends which agencies can use to inform their responses in emergency situations. Differential privacy allows organizations to obtain this kind of location data and mobility insights in a hyper-safe, anonymous environment.

Contact us for more information on how the Spectus Data Clean Room can serve as an integral privacy tool for any organization's needs.

spectus™

Data. In sight.

US Office
45 West 27th Street
3rd floor
New York, NY 10001

www.spectus.ai