replicahq.com
16 Maiden Lane
San Francisco, CA 94105

Replica Uncertainty Analysis

Version: 04/17/2020

This document serves to highlight responsible use of Replica data and products through communicating strengths and weaknesses of our model.

What is Replica?

> Replica is a synthetically-generated representation of the activities and movement of residents, visitors and commercial vehicle fleets in a metropolitan area for a representative week during a given season. Replica represents movement by combining data from three primary types of data: public use population census data, proprietary locational data from telecommunications and other IT infrastructure in the region, and field observations data from customer public agencies (ground truth). The purpose of this document is to provide a high-level overview of Replica's data processing methods and statistical inference systems in order to evaluate quality, accuracy, privacy and data security implications.

Fundamental sources of certainty in Replica

Each trip in Replica is modeled to reflect a "real" trip that could take place on a given day in a region.

Larger sample sizes

The Replica model is most statistically accurate when viewed in aggregate and therefore, with larger sample sizes our data becomes more accurate.

This sits in line with general statistical and surveying principles: *with a larger sample size, we can be more sure we are looking at a representative picture rather than one affected by outliers*. This principle holds true in paper surveys, statistics, and other realms.

Principle applied:

- Replica will be most accurate with larger sample sizes. For example, we are confident in saying that there are 8.16 million trips in Portland on a typical Thursday in Winter 2019, but filtering to 10,000 trips would be looking only at 0.12% of modeled trips and certainty should scale down appropriately.
- Filtering by many particular trip characteristics is similarly going to reduce the sample size and risk the confidence afforded by aggregation. For example, the number of shopping trips by Black residents from a particular tract during rush hour will be less accurate than looking at trips with fewer filters.

Data inputs for Replica

As explained in our Methodologies document, our model is rooted in various forms of input data. Some of that data is provided directly by our clients, like sensor data, and some is consumed from third-parties. Collectively, these sources allow us to know that our model is accurate.

Each source of data will have its strengths and weaknesses. Using multiple sources of data allows us to make a more accurate model. For example a misfiring sensor (which appears as an outlier) is not useful to include as-is in our model, similar to how it would not be useful when manually reviewing its provided counts. Additionally, sensor information is not available on all streets and paths. So while we use data sources and sensor information to calibrate our model, there are further optimizations and gaps to fill. It's Replica's combination of data sources and statistics/technology that provides an accurate picture of transportation in a region.

Spatial data input quality

In order to present accurate travel prediction at scale, Replica necessarily relies on input sources that inform how our modeled trips can move. Spatial data includes road and bike networks from OpenStreetMap, transit lines and schedules from agency-provided GTFS, and other data points like business locations and schools.

When those input sources are inaccurate or incomplete, this directly impacts our model's outputs and accuracy. For example, inaccurate turn restrictions in input data would cause inaccuracies as we route personas. Another example is missing GTFS bus lines, stops, connections, and/or weekday/weekend variations, which would cause incorrect transit routing.

replicahq.com
16 Maiden Lane
San Francisco, CA 94105

Modeling minors:

Our data suppliers are legally prohibited from collecting the data of minors. This makes minors an example of a category of residents that are necessarily simplified due to our data inputs. Minors are assigned a school based on their age and the school's proximity to home. Overall enrollment is based on publicly available enrollment data. Home locations are based on census data. We do not take account of additional factors that may contribute to school choice.

Modeling uncalibrated locations:

As roads increase in size and importance, our trip estimates grow in confidence accordingly. In the other direction, our modeled trips are less accurate for smaller (usually residential) roads. These numbers will–in general–be reflective of observed daily use, but there will be occasional inconsistencies between our model and reality. For example, Replica drivers may choose to turn onto a side road which saves a few minutes of commuting time, whereas an actual driver may prefer to continue straight, avoiding any detours.

Additionally, sensors are more likely to be placed on large roads, reducing the input data available to calibrate against for smaller roads and pathways.

Modeling transit boardings and alightings:

With our current capabilities, Replica does not calibrate transit boardings and alightings to customer data. The ridership for each line will generally be reliable, and in general, our synthetic population will tend to board and alight at realistic stops. However, actual people may choose to walk slightly farther to a more comfortable bus stop, or one which lowers the trip fare.

Principle applied:

- Use extra caution when looking at the above categories affected by data sources. To gain certaintanty, focus on aggregate trends.

replicahq.com
16 Maiden Lane
San Francisco, CA 94105

Examples of strong Replica use cases

Investigating Transit Ridership

Replica can be used to understand transit ridership by demographic and purpose. Examples:

- With Replica one can find out what demographics are shared by riders of a particular line or sets of lines and then go on to contrast that with other characteristics or transportation options for those residents.
- With Replica you are able to focus on which residents live near or far from public transportation options.
- Findings can be further understood through cross-seasonal comparisons or by day-of-the-week comparisons.

Analyzing transportation trends

Replica can also be used to look for active travel promotion opportunities. Examples:

- Evaluating short auto trips to identify where there is opportunity to promote walking or biking alternatives.
- Understanding and comparing residents' VMTs across a region.
- Identifying workplace transportation trends for major employers in an area to see where employees are coming from and how they are getting to the selected area.

Project planning

Replica can provide a baseline for project planning, highlighting areas of interest. Examples:

- When considering implementing a project like congestion pricing or adding a new transit stop, Replica can highlight where trip makers are travelling from and their associated demographics–including people coming from the buffer zone of the Replica.

Assisting public works

Replica can help allocate resources for infrastructure improvements. Examples:

- Replica can provide bridge counts so an agency can marry this data with their own to better prioritize infrastructure improvements and investments.

Understanding Resident Spending Habits

Replica can provide context on where and what residents are spending money on. Examples:

- Replica can provide trip destination and trip purpose of residents from a particular geographic area within a region so agencies can determine where potential sales tax dollars are going.

- Replica can provide a baseline and seasonal refreshes to determine if particular investments cause increases in traffic and/or spending.

Examples of cautionary Replica use cases

Has a new trail increased walking and biking?

Use caution. The street network (including bike trails and lanes) is based on OpenStreetMap data for the relevant year. The network is not updated on a seasonal basis however we do make specific updates to take account of significant long term closures or new openings. Temporary closures of minor streets may not be reflected in the model.

Additionally, due to the limited coverage of mobile location data, Replica may not detect behavior changes by small numbers of people. Recreational trips such as walking a dog or jogging are not currently included in Replica, and trips made on the trail may not be counted as primary mode walking trips. Due to limited ground truth data for walking and biking, these numbers are uncalibrated and should be used with caution.

Lastly the algorithm Replica uses to route bicycle trips does not fully take account of factors such as elevation or perceived safety.

How many people are driving on a minor street?

Use caution. As discussed in the "Modeling uncalibrated locations" section, although trip counts on all streets in Replica are observable, smaller streets are more likely to contain outlier data and are not intended to replace traffic counts. Replica data may still provide insights and trends.

How did a new transit schedule affect ridership?

***Use caution*.** The public transit network is based on GTFS data for the relevant season. We are reliant on transit agencies to ensure that changes in transit networks are accurately reflected in published GTFS data. Additionally, GTFS data provides scheduled transit times, not real time performance.

Examples of not advisable Replica use cases

How do children get to school?

Not advisable. Our data suppliers are legally prohibited from collecting the data of minors. This makes minors an example of a category of residents that are necessarily simplified due to our data inputs. Minors are assigned a school based on their age and the school's proximity to home. Overall enrollment is based on publicly available enrollment data. Home locations are based on census data. We do not take account of additional factors that may contribute to school choice.

How does the availability of parking affect private auto usage?

Not advisable. In Replica drivers park at their destination. We do not currently account for the availability of parking or route people to nearby parking lots.

How do people drive, bicycle, or scooter to public transit?

Not advisable. We do not model biking or driving to/from public transit. We do not model scooters.

replicahq.com
16 Maiden Lane
San Francisco, CA 94105

Examples of unsupported Replica use cases

Where do people work from home?

> ***Not supported***. We don't currently model working from home. These people will stay home, but will not be distinguishable from others who stay home.

How much congestion is caused by a particular event? (e.g. a sporting event at an arena)

> ***Not supported*.** Replica is a model of the movements on a typical day in a three month period, as such congestion due to special events would not be reflected in the model.