Azira™

Mobile Location Data Normalization:

An Azira Primer

Los Angeles | Paris | Bangalore | Singapore | Sydney | Tokyo

www.azira.com

Introduction

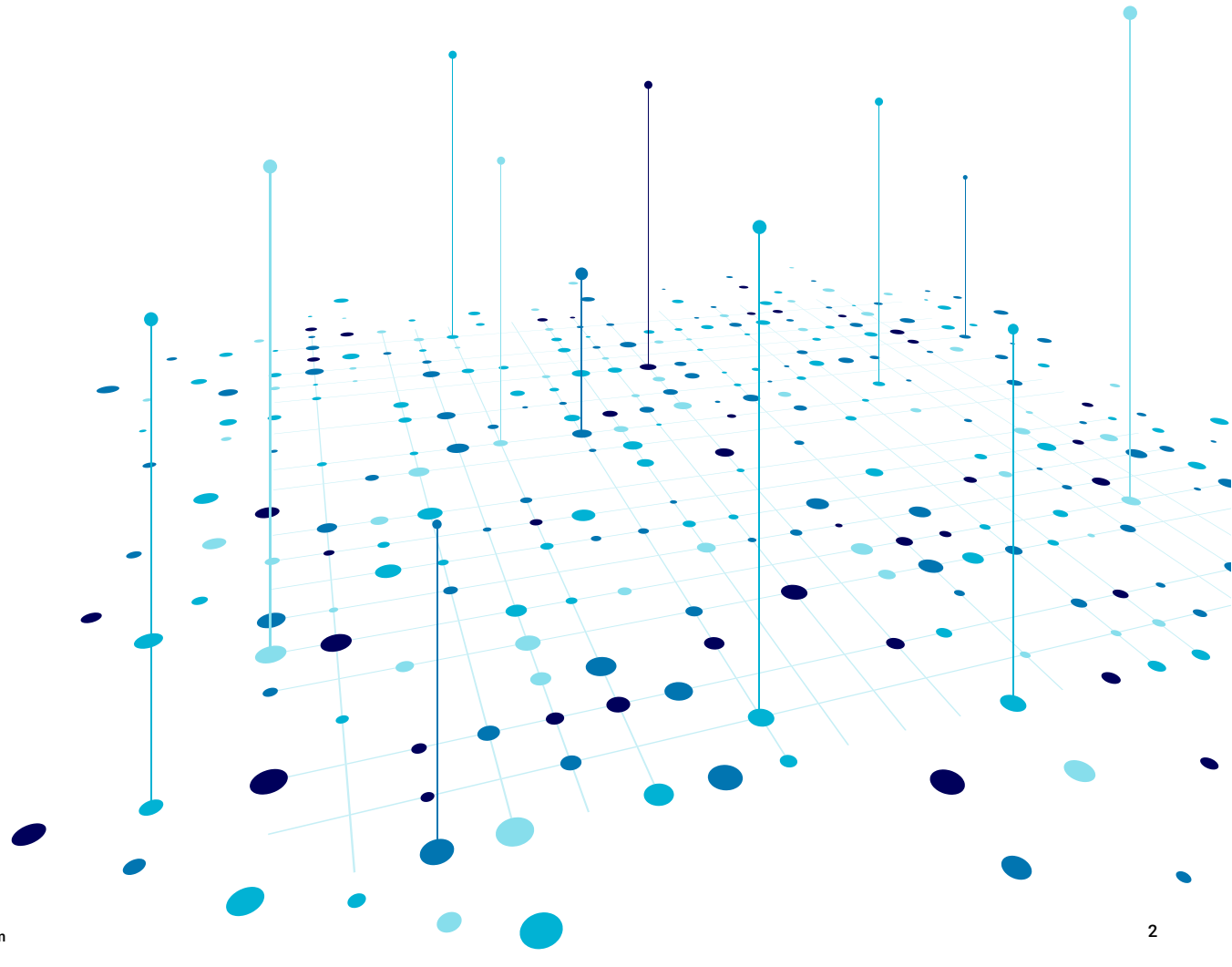
The volume of mobile data available on the market at any given time can vary greatly, depending on a number of factors. Although Azira applies rigorous quality screening to clean and structure the data we sell, these efforts alone do not offset underlying volatility in the ecosystem. For time-series analysis specifically, customers working with our data are encouraged to apply data normalization methods.

The objective of location data normalization can vary but primarily focuses on three areas:

- Data smoothing minimizes volatility, reduces spikes and fluctuations and generally makes the dataset appear more consistent.
- Data scaling allows you to compare variables on an equal footing, and
- Data baselining, which defines a baseline for relative comparisons and trend identification.

Some organizations conflate smoothing with normalization because the end result simply looks better, so it's easy enough to declare victory. However, to achieve a better analytic result, truly effective normalization requires a combination of approaches to ensure a more accurate reflection of real-life trends.

In this white paper, we will explain and evaluate the benefits and limitations of common normalization techniques. We will share approaches to remove false variability caused by a unique set of challenges associated with the mobile data collection process and demonstrate a model that eliminates volume fluctuations and extreme outliers.



Background: The Volatility of Human Movement

Azira's data is generally classified as a spatial time series. Each record contains a device identifier, precise coordinates and an exact timestamp. This kind of spatiotemporal data uniquely requires normalization to gain accurate and relevant downstream consumer insights or to feed ML models with confidence, for the following reasons:

Data trends may deviate from observed ground truth: data collection limitations

The visitor trends for a given store or place of interest may not identically reflect the ground truth because Azira only collects a sample of the actual visitors, depending on the specific apps installed on users' mobile devices, and the users' mobile permissions for those apps. For "always on" SDKs to collect the relevant datasets, users need to opt-in to foreground sharing and turn on their location services while in a specific store or point of interest. There are certain locations that cannot be reached by cell towers, places and events where mobile devices are not permitted, and other reasons that can affect the sample size. Also, not every app or SDK collects and transmits locations at a high frequency; if a POI tends to have shorter visitations (for example, a take-out-only restaurant), it may not always be captured by an SDK.

Even within a specific high-quality data source, there can be reasons for anomalous fluctuation. For instance, solar flare activity can create inaccuracies in GPS information. Or a base station on a cell tower may be compromised due to a parts failure, disrupting service.

Ground truth itself demonstrates high variance

Seasonality impacts physical world behavior. For example, low footfalls will be seen in open public places during a winter storm. A holiday or event may "artificially" elevate or decrease observations.

In addition, factors such as population density, customer engagement time as influenced by local cultural norms, and a store's size, location or brand popularity may impact the observed footfalls. For example, footfalls vary between a Starbucks store in Manhattan vs. a Starbucks store in rural North Dakota; footfalls vary between a large-format IKEA vs a Claire's store in a shopping mall in the same city; footfalls vary between a convenient McDonald's vs a "destination" fine dining restaurant.

Data fluctuates because of source variability: data supply limitations

Azira's consumer behavior data captures a sample/representation of the actual visitors for a given store or point of interest. This capture rate can vary for multiple reasons:

- → OS changes, regulatory changes, or publishers modifying their offerings and monetization strategies.
- → Changes in the app ecosystem can be a significant and unexpected factor in fluctuations. For example, in September 2023, a change in Google Play store API requirements affected a large number of apps, temporarily decreasing data availability.
- → Data sourcing is a complex ecosystem. Just as Azira continuously evaluates and adjusts the data source partners we work with, our provider partners also change the publishers they work with, affecting data depth and breadth. When we find new interesting sources of data, we add them. By that same token, we also turn off data that no longer adds to our universe. We may also switch data providers:
  - To limit duplication of devices and points as much as possible
  - If the overall quality of the events drops
  - For contractual or legal reasons.

The high variance of Azira's data can result in skewed consumer insights and poor ML model performance. While there is no perfect solution for eliminating noise and improving the overall accuracy of the data, the sections below outline some approaches that can help pave a path toward directionally correct insights.

Approaches to Location Data Normalization

Before we step through one possible approach, let's review common requirements for and approaches to normalization, how they support different objectives, and what some limitations might be.

Data cleansing is a prerequisite for normalization, mainly designed to remove invalid or fraudulent data, as well as synthesized data that has been added by some upstream sources to artificially increase data volumes (for more information, see the blog post, Data Anomalies 101). All Azira data goes through two cleansing phases. The first runs the data through a set of rules to determine which data points are valid, without taking any additional context into consideration. The second cleansing phase incorporates statistical methods to remove anomalies, which can potentially eliminate some legitimate data points.

Data smoothing removes outliers from a dataset, resulting in a less visually noisy output (i.e., fewer spikes). Simple smoothing methods, such as moving averages, are not nuanced and run the risk of deleting some of the signal with the noise. As a result, visualizations may look superficially "nicer," with fewer spikes, but the overall quality of the data (i.e., how accurately it represents reality) may not see significant improvement.

Selective data sampling filters out devices (and/or data points) by using specific qualitative attributes. By carefully selecting certain devices based on their historical activity, or intentionally matching a certain distribution for a given attribute, the overall trends of the data can be bounded.

- **Sampling by device frequency**
  Initially, when working on our own normalization models, Azira started by sampling devices with higher observability. This approach makes each device more credible (assuming the data points are not anomalous), and the higher volumes of historical data points make the data more appropriate for some machine learning applications. However, depending on the thresholds used for filtering, this approach can also, paradoxically, significantly reduce the amount of data. More recently, we have tried to replicate the "natural" distribution of devices, which has proven more useful for interpolation and prediction.

  In both cases, the resulting analyses tended to highlight whatever biases remained in the data after the filtering, so this approach is only recommended if executed in conjunction with additional data.

- **Sampling by demography/geography**
  Similarly to the device frequency filtering, the panel can be modeled to match a particular demographic or geographic distribution. If (modeled) home locations are being used to tie devices to demographic/geographic data, then the methodology for determining those locations must be reasonably robust. As with the frequency approach, regardless of how the panel is generated, ground truth is critical for validating results across a large number of POIs. Otherwise, any analytics performed on that data will skew towards trends that exist in the sampled data but are not necessarily reflective of reality.

Baselining uses a secondary data source (often at some level of aggregation) where the relative impact of the noise is better understood. For example, if there were fluctuations in the mobile location data for a nearby residential area that is expected to have a reasonably constant population, those fluctuations could be deducted (in a relative manner) from the data for a POI, assuming the causes of those fluctuations are the same. At a country or municipality level, Azira's Data Volume Report can be used for the same application, to remove fluctuations due to data collection and source variability discussed above. The best data for baselining is ground truth, as this allows for more accurate normalization for POIs in the same category (for example, baselining with Starbucks stores for a given geo may yield better results for normalizing Peet's store data).

Data Aggregation by time and area is not technically a distinct normalization methodology, but it is a suggested best practice for executing any of the normalization techniques described above. Given the nature of location data, it is not always feasible to normalize and estimate by day for a given polygon. In fact, many of our customers have validated that Azira's data is most accurate when aggregated by month or year, or across multiple POIs or geographic regions (zip codes, DMAs).

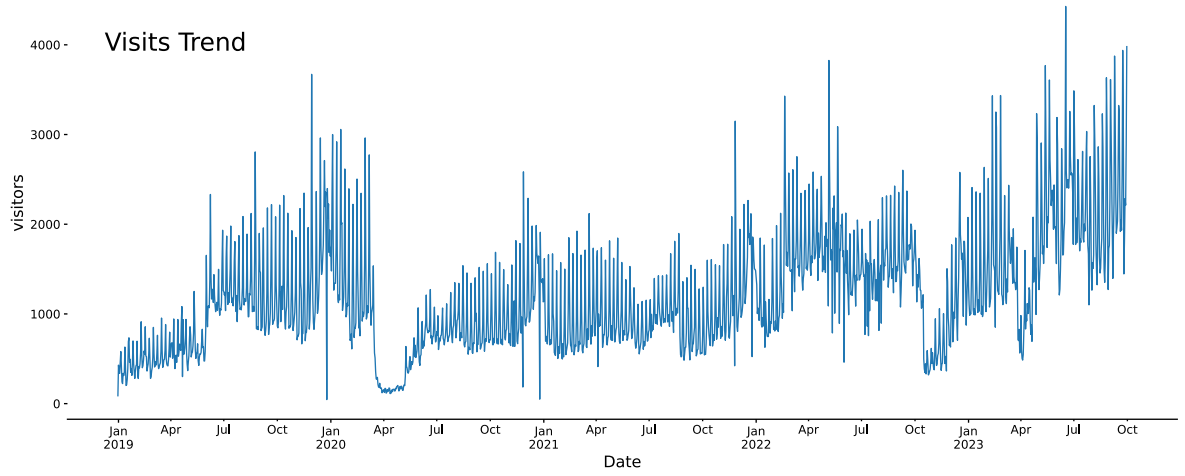
Historically, Azira has experimented with all of these methods. Our recommended approach combines both baselining and selective sampling, but again, in the absence of ground truth, there are severe limitations in building and validating normalization models.

For further consideration

- Normalization is recommended for visitation count reports such as pin or zero-point reports. Azira does not normalize this data by default.
- Normalization goals and methods can be specific to a customer, brand, category, point of interest, data density and relevant use case.
- Not all consumer behavior data sets need to be normalized, and not all customers will want to or should normalize data the same way.
- Data sets such as CEL (Common Evening Location), CDL (Common Daytime Location) or Pathing need not be normalized because they are derived insights.

Normalization Example

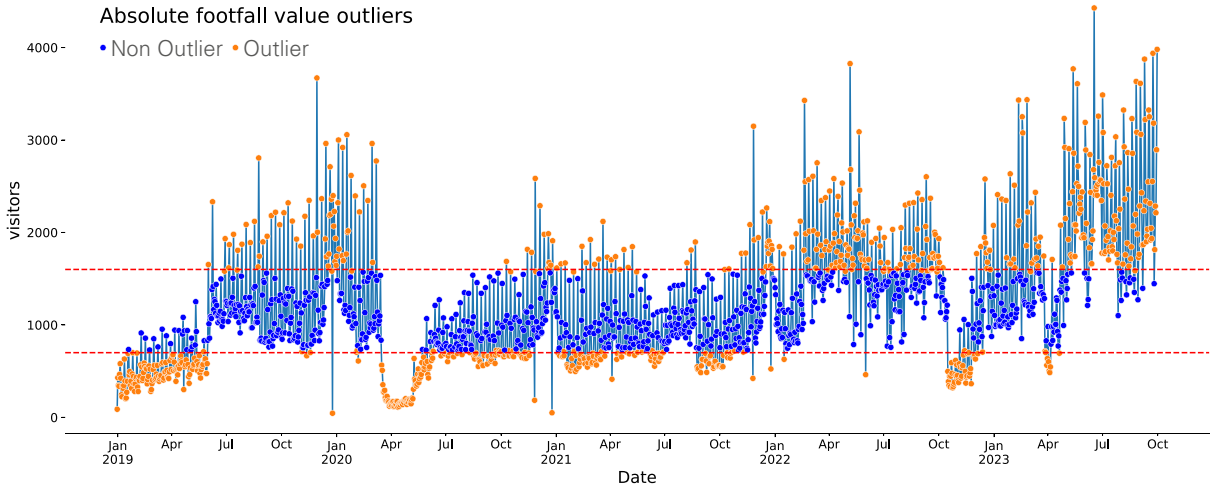
In this example, we examine the footfall trend of a shopping mall in California. The data shows fluctuations due to many of the factors discussed above, including but not limited to, changes in mobile usage patterns in the shopping mall and variations in the incoming data volume from Azira's partners. We do see an expected drop at the onset of the COVID-19 pandemic, but there is an increase in daily fluctuations starting around April of 2022.



The following steps outline a normalization methodology that creates a baseline using the data itself. Note that if the data quality is relatively high and the trends are accurate, this type of normalization is sufficient to remove noise. However, if the data is extremely biased and/or there are large amounts of noise, additional data (from non-mobile sources) may be required to establish a more accurate baseline (see Approaches to Location Data Normalization above).

Methodology

- **Step 1:** Using the Azira platform, extract daily footfall values for a specific location. This methodology works best when applied to multiple years of data so that we can establish a baseline that accounts for seasonality.
- **Step 2:** Determine the outliers in the data. As a starting point, we consider the top and bottom quartiles as outliers so that we only keep the middle 50% of the data as is. This entire process can be repeated later using different thresholds, depending on other statistical factors such as the variance of the data.



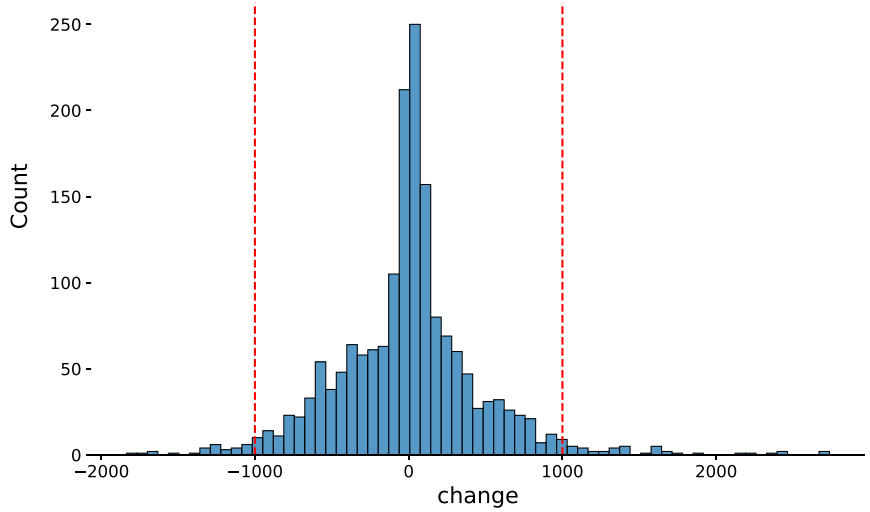
$$\text{Filtered values} = \{x_i \text{ if } x_i \in [l,h]\}$$

l = lower percentile
h = higher percentile
x_i = data points

If, for a given footfall time series, the minimum and maximum values of the 50th percentile are 50 and 150, respectively, then for a sample data series of [10, 50, 60, 109, 170, 145], the outliers are 10 and 170.

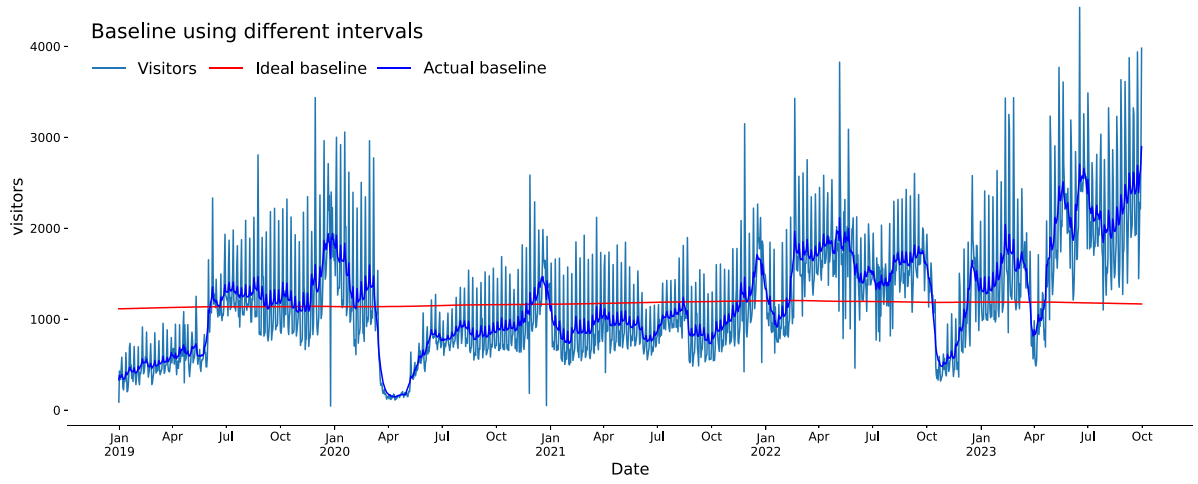
- **Step 3:** Next, we replace the outliers identified in the previous step with non-outlier values. One way to achieve this is to employ forward and backward sampling from the sorted non-outlier values on time/date and calculate a weighted average to replace the outlier value. For instance, if 10 and 170 are identified as outliers, an extremely simple version of this approach would be to use the next value in the sorted sequence, resulting in replacement values of 50 and 145, respectively. If there are specific dates (for example, Black Friday or Christmas Day) that are known to have exceptional values, these can be omitted from this step and added back into the time series at the end of the final step.
- **Step 4:** Create daily change data by subtracting the consecutive aggregated footfall value from the previous day's footfall value.
  - Remove change outliers using the standard deviation method, unless there are specific data points that should be preserved (see Step 3).
  - Fill in the outlier value using the average change value or forward/backward sampling for neighbors.
  - Keep the change sign the same as in the dataset. This step is necessary to remove the set of outliers due to changes in the daily absolute footfall value.
  - Ignore the set of change values that fall too far away from the center of the distribution.

- Below is the example distribution of the change values and the cutoff values for removing the outliers based on standard deviation. For example, if the second standard deviation is 100 for a given set of change values, then all values that are more than 100 or less than 100 will be marked as outliers, and their values will be recomputed using an average change value.



→ **Step 5:** Create a baseline using moving averages. We start by creating two series of moving averages, one using a longer period (e.g., 365 days) and the other using a shorter period (e.g., seven days). Unless the overall longer-term trend of the original data shows a significant increase or decrease, the moving average using the longer period should be relatively flat. In the chart below, the red line indicates the moving average with the longer period, and the dark blue line indicates the moving average with the shorter period.

$$\textit{Simple moving average} = \frac{(x_1 + x_2 + x_3 + \ldots + x_n)}{n}$$

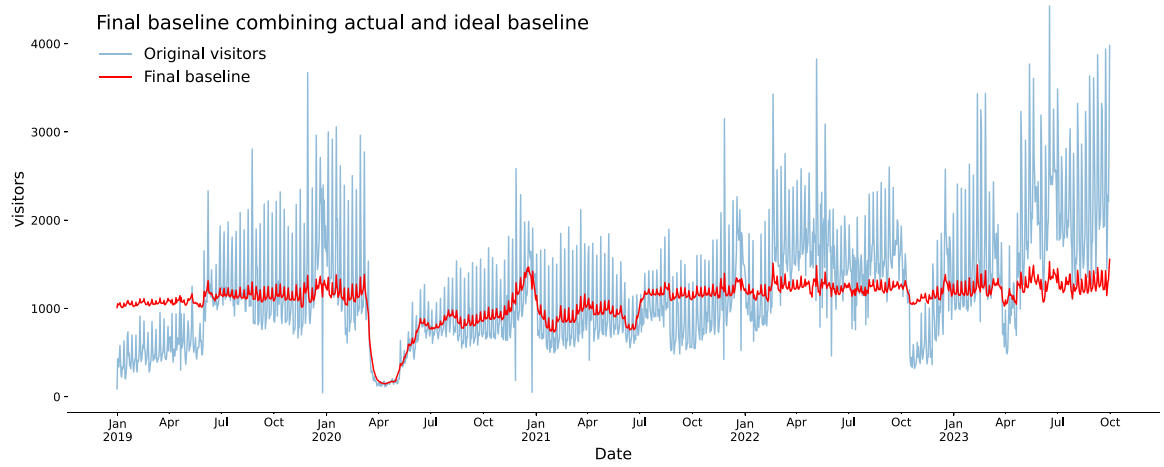


→ **Step 6:** Combine the baseline from Step 5 with the daily change data from Step 4 to create the final normalized trend. This step is essential to include the effect of natural data volume changes and also stabilize data volume across a given timeframe. The weighting can be adjusted depending on how much emphasis should be placed on the longer-term trend compared to the shorter-term trend (i.e., the periods used for calculating the moving averages in Step 5).

$$b3 = \frac{(w1 * b1 + w2 * b2)}{(w1 + w2)}$$

b3 = final baseline
b1 = first baseline
b2 = second baseline

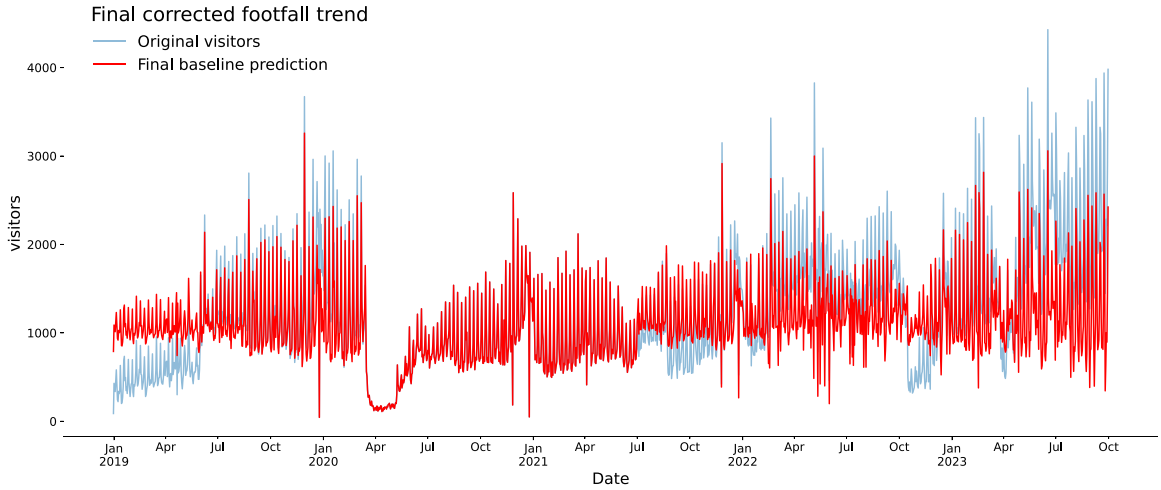
w1 = weight assigned to b1
w2 = weight assigned to b2



→ **Step 7:** Combine the final baseline from Step 6 with the daily changes calculated in Step 4 using the following formula:

$$corrected\ footfall = b3 + \Delta x + bias$$

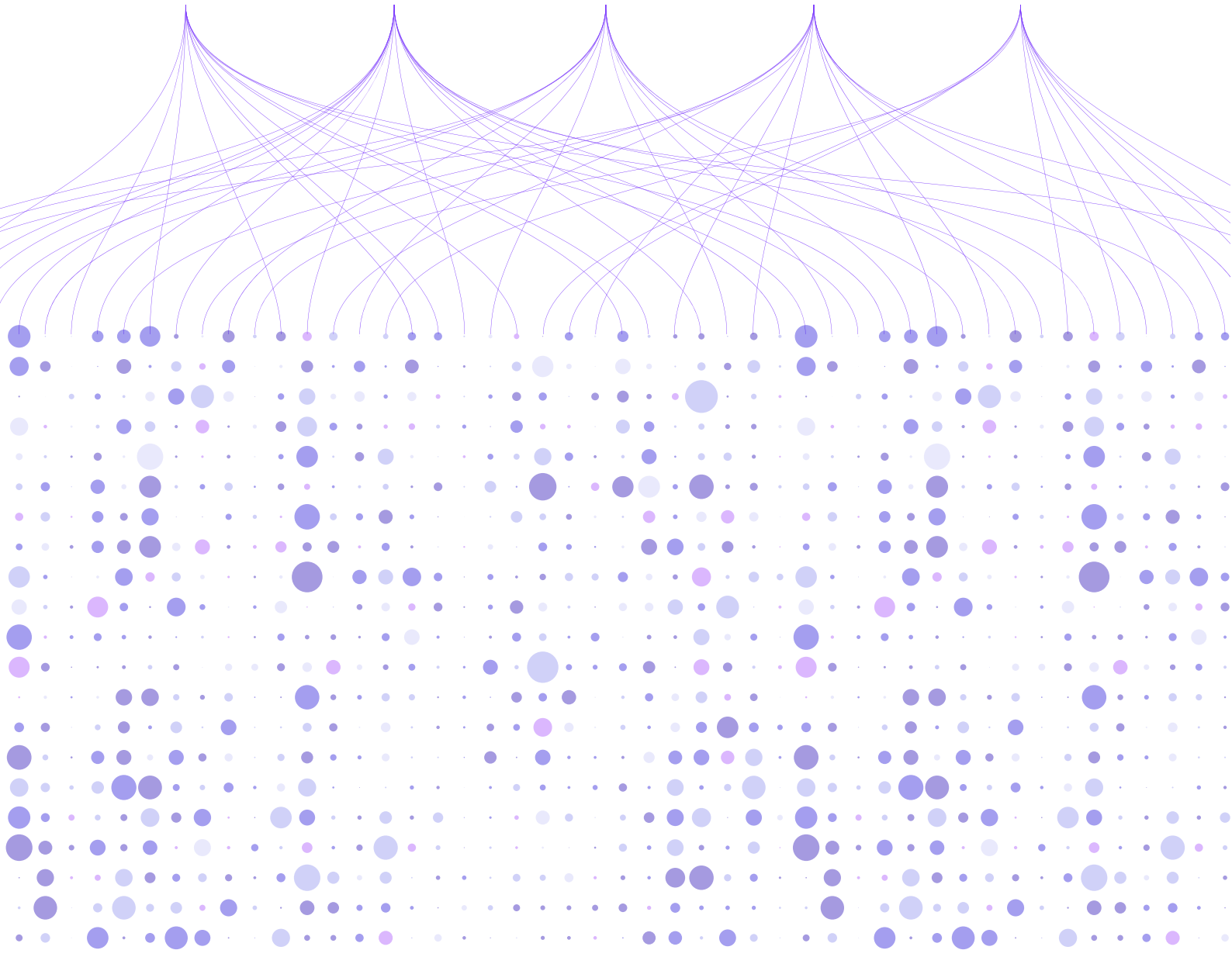
b3 = final baseline
Δx = daily footfall change
bias = value to avoid negative final value of footfall



Conclusion

The output of any normalization exercise should be carefully reviewed against what you already know about the trends you are evaluating. Does it pass the “sniff test”? Does it look right? For example, how do results compare pre- and post-COVID? Do they reflect seasonal trends (Black Friday, weekend vs. weekday), etc.? As mentioned multiple times in this paper, normalized results should be validated with ground truth whenever possible.

Mastering data normalization requires understanding the data's nature and nuances, selecting appropriate techniques, and evaluating the output data variances. Embracing these best practices and iteratively implementing multiple techniques and ongoing process improvements will result in high-quality and reliable data, feeding robust machine-learning models and ensuring accurate insights to drive your business forward.



Azira™

About Azira

Azira LLC, a global Consumer Insights platform, helps marketing and operational leaders improve their effectiveness with actionable intelligence to drive business results. Its mission is to create a more relevant world where brands are empowered to reach and build relationships with their consumers. With a profound commitment to partnership, trust and transparency, combined with decades of expertise in consumer behavioral analytics, Azira delivers innovative marketing solutions to curate audiences, activate omnichannel campaigns, and understand footfall attribution. It also provides operational insights for use cases such as site selection, trade area analysis, competitive intelligence and more. Azira serves enterprises in retail, hospitality, travel, real estate, financial services and media. A global company, Azira is headquartered in Los Angeles with offices in Paris, Bangalore, Singapore, Sydney, and Tokyo. To learn more, please visit https://azira.com.