Community support intensity in urban settings: an empirical, place based model

Nitish Verma[1]

[1]Essence Networks Limited

March 11, 2023

Abstract

Virtual street audits are increasingly used by social service providers, community leaders and social workers to support place-based decisions in urban settings. Analytical tools can potentially augment decision making by making relevant information accessible in context to place-based interventions aimed at improving community wellbeing. I investigated if it was possible to discover a quantitative stochastic model that can be a reasonable fit with empirical data about the relationship between physical urban locations and the volumes of public services supplied to those locations. I analysed the dataset of San Jose's Police Department call centre data spanning 10 years. I found that the Borel-Tanner distribution is a reasonable fit in 7 out of 10 trials conducted in this study. Relatively few urban sites were attributed to a large proportion of service volumes. There was high spatial concentration within 'high needs' sites. A significant proportion of these sites persist over the 10 year period. I plotted these sites on a map to demonstrate their applicability to virtual audit applications.

Introduction

Virtual streets audits are techniques that enable urban planners, researchers and professional social workers make informed decisions about local communities by the use of digital technologies such as "Google Street View" (Badland et al., 2010). The purpose of such audits is to better support better wellbeing outcomes for urban communities. I suggest that virtual street auditors will be better served with more context-rich information to support their assessments. The objective of this paper is to present the research done to (a) examine the suitability of government open data to provide this additional context. (b) study the quantitative model relationship between physical urban areas and the volume of service delivery provided to those location areas. (c) Visualise the spatial characteristics and trends in context to the services being studied. (d) to conduct analysis consistent with reasonable public expectations of the privacy of individuals and groups of individuals.

Specifically, I examine the model relationship

$$P \rightarrow S_{\mathrm{p}}$$

between $\mathbf{P}$, the location and the $\mathbf{S_p}$, quantity of service delivered to that location.

I formulate the primary hypothesis (**H1**) that there exists a stochastic quantitative model that reliably characterises the relationship between $\mathbf{P}$ and $\mathbf{S}$. My assumption that each SJPD visit is regarded as a discrete and independent event. Therefore, I limit my research to the family of probabilistic discrete distributions. I suggest that a ChiSquare test result with P ≥ 0.05 would be a reasonable confirmation of my primary hypothesis.

In this paper, I examine over 6 million records of publicly available government open data: the San Jose Police Department's call center data, sourced from the San Jose City ("Police-Calls-for-Service - San Jose CA Open Data Portal", n.d.) , over a ten year period, starting May 2011 and ending May 2022. The dataset is updated daily, as a time series, with a latency of 1 day.

I define the quantity of service provided on the site as the number of visits that SJPD physically makes to the location address recorded in the call record. My rationale for this is that, this way, there is some reasonable assurance that that the final determination of the outcome of the visit (recorded as a 'final disposition') will be governed by the Department's internal policy.

I define the class of service - "community support service" - as the service requested by the public to support them in the context of "Family Disturbance" and those that are not assessed by the SJPD as resulting in a decision to Arrest, Cite or perform any other serious criminal justice intervention. In other words, I assume that in cases where the SJPD has decided not to take adverse legal action, it is essentially acting in a community support role in predominantly family related matters. I take the disposition of "R - Report Taken" as the threshold at which to include the visit for study. To be clear, I include disposition statuses of "O- Unfounded Event", "G - Gone on Arrival" in the study because these events suggest that they are not trivial and that these disposition codes most probably reflect SJPD administration policy - not necessarily the family's context in question.

Methods

Data preparation

I acquired the data from the server and performed consistency tests. I fixed data format quality issues in relation to timestamps.

Data Selection

I selected data for 10 annual periods, starting from May 13, 2011 to May, 12, 2021.

Data extraction involved the subsetting of the primary dataset, by "CALL_TYPE". Only "DOMESTIC DISTURBANCE" calls were selected. I further subsetted data to remove all entries that related to serious adverse action by SJPD. Instead, I selected those records where I believe SJPD has made a formal determination after actually visiting the site.

N	No report required; dispatch record only
G	Gone on Arrival/unable to locate
R	Report taken
H	Courtesy Service/Citizen or agency assist
O	Supplemental report taken
U	Unfounded event
T	Turned over To (TOT)
NR	No Response
F	Field Interview (F.I.) Completed
P	Prior case, follow-up activity only

Table 1: The police call disposition codes that are included in calculating counts of services provided to a location site.

Constructing the model variables

A location, (**P**) can be described as a bivariate coordinate expression (x,y), but for this purpose, **P** is operationalised as the approximate city address. The number of SJPD visits to the location **P** is the dependent variable ($\mathbf{S}_{\mathrm{p}}$) is operationalised as a non-negative integer variable.

Each police call record in meant to contain an address field. I record each unique address as a unit **P** variable, which is sorted by $\mathbf{S}_{\mathrm{p}}$. Th highest ordinal number is Max (S_{p}) within the time window being analysed.

To generate $\mathbf{S}_{\mathrm{p,}}$ I use the OFFENSE_DATE to determine the date at which the visit to the site occured.

Exploratory data analysis

Exploratory data analysis involved (a) Time series visualisation of total amount of services delivered per day ($\mathbf{S}_{\mathrm{d}}$) across the full time range. (b) For each annual period, a simple X-Y plot of locations of **P** against $\mathbf{S}_{\mathrm{p}}$. (c) Generating descriptive statistics of the variables **P** and $\mathbf{S_p}$.

Model discovery and fitting

I attempted to find the appropriate stochastic model iteratively fitting the service volume data to the distribution and observing result of the ChiSquare test.

Longitudinal analysis

I performed longitudinal data analysis, yearwise, with a focus on locations. I selected the top "high needs" locations that were common in each of the ten years.

Visualisation

I geocoded the locations, by (unique) street addresses and plotted the result geospatially.

Privacy and Ethics

Address locations are sensitive, personal identifiable information, especially in context to adverse family related events. I assessed the data set and research plan for privacy risks. The source system has reasonable privacy protections because the address locations are expressed in ranges spanning a city block. In processing the data, I do not present or report any specific location at any stage as I have assessed that the actual location is not important to the study objectives. Geospatial plots produced in this paper does not present the basemaps which therefore cannot link the locations to street geometry.

Results

Exploratory Analysis:

The results of exploratory data analysis is shown below:

1. Time series of observation(S)

The time series plot of total counts of services delivered per day across all places, ($\mathbf{S_d}$) over and 11 year period is shown in the figure below. A rolling 10 day mean of $\mathbf{S_d}$ is also generated. The sharp jump in service volumes in the last year (May 13, 2021 to May 12, 2022) was unexpected. As discussed in the Limitations section in this paper, I dropped the 1 year period and used data until May 12, 2021.

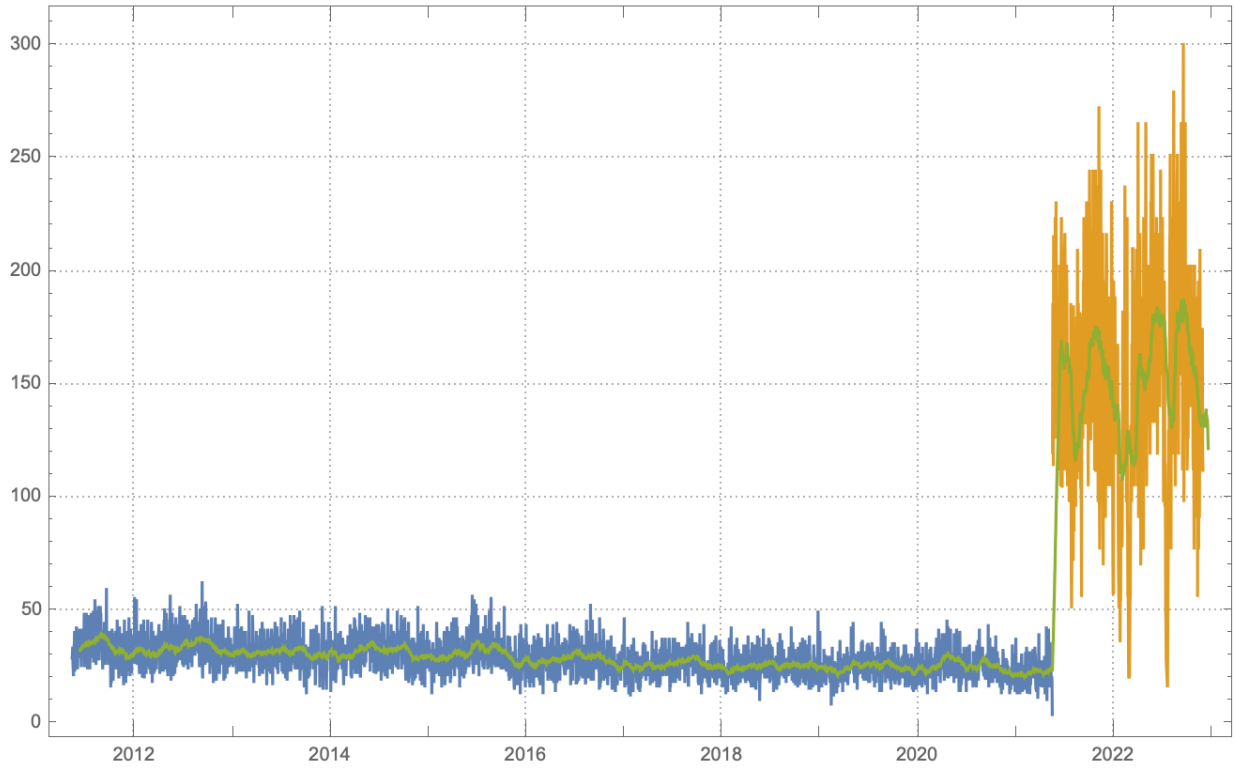

Figure 1: Time series observations - total counts of service visits to all sites, per day (May 2011 to May 2022)

Time Period	N	Mean	SD	Variance
10 Years	3653	28.71	7.58	57.38

Table 2: Descriptive statistics of S_d over 10 years

2. Model variables, $\mathbf{P}$ and $\mathbf{S}_{\mathrm{p}}$

Over the 10 year period $i \in \{1..10\}$, I found 20,950 unique addresses (n($\mathbf{P}$)).

The total service counts across the same period was

$\sum_{i=1}^{10} S_{p,i} = 104{,}861$

I generated X-Y plots of unique locations ($\mathbf{P}$) against total counts of service visits to those locations ($\mathbf{S}_{\mathrm{p,i}}$), over annual periods. These are shown below. The X-axis (P) is plotted in descending order of ($\mathbf{S}_{\mathrm{p,i}}$).

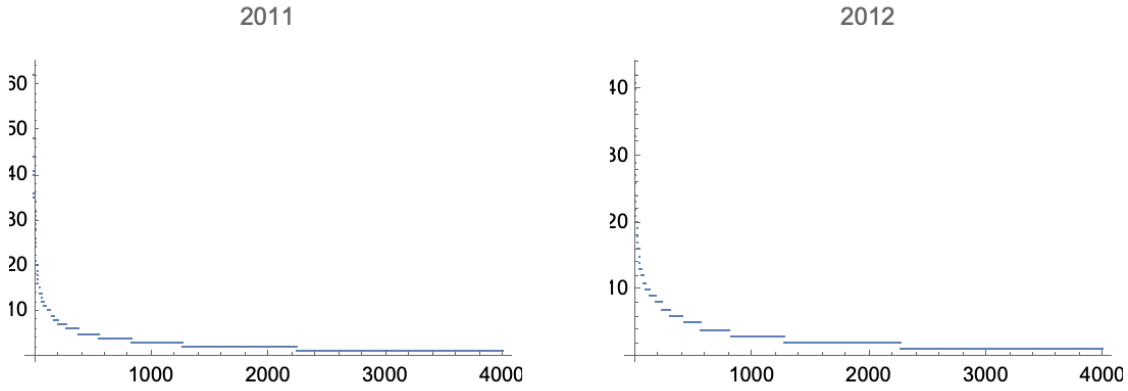

Figure 2: X-Y plot for 2011, 2012 (X = P, Y =$\mathrm{S_p}$)

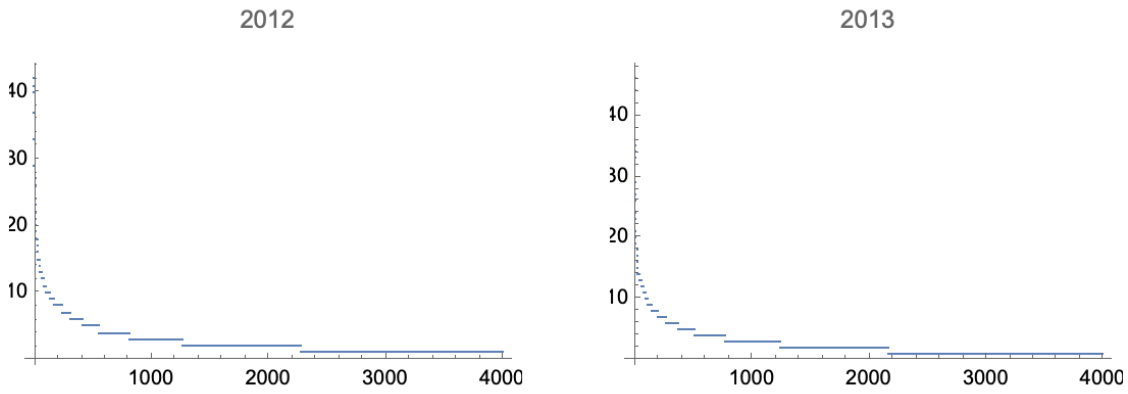

Figure 3: This is aX-Y plot for 2012, 2013 (X = P, Y =$\mathrm{S_p}$)

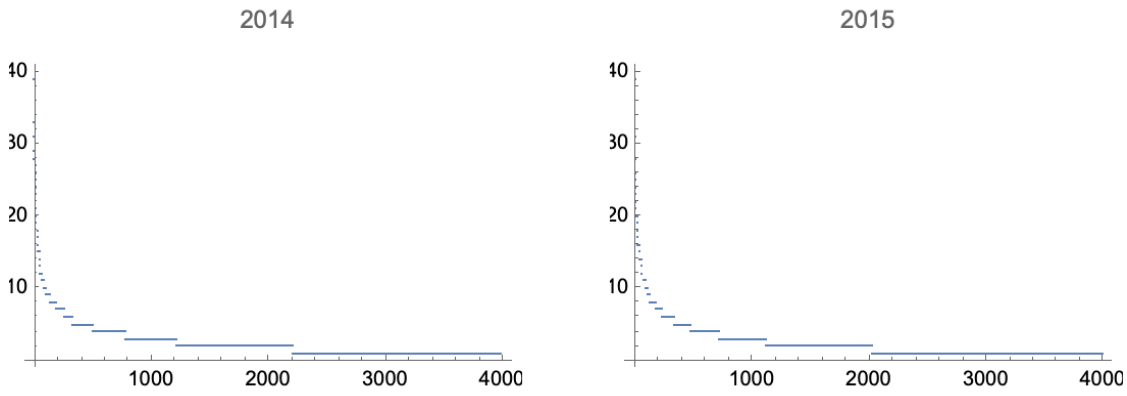

Figure 4: X-Y plot for 2013, 2014 (X = P, Y =S_p)

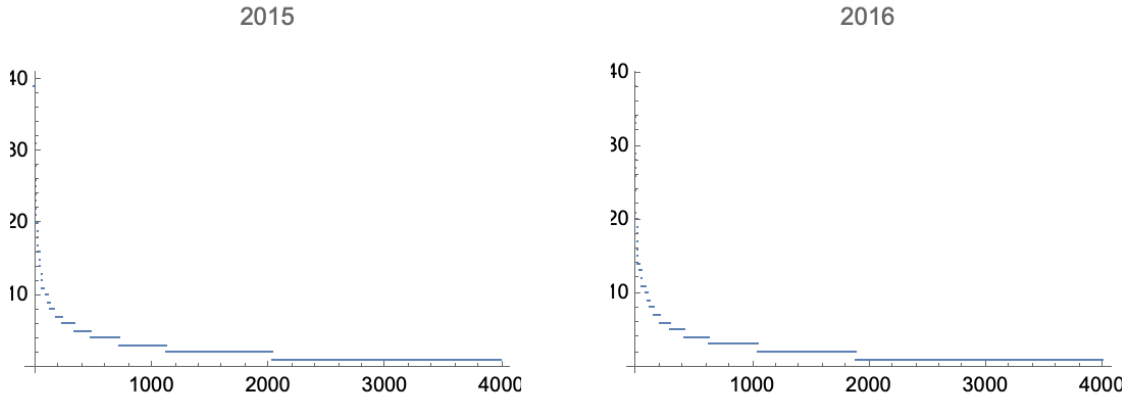

Figure 5: X-Y plot for 2015, 2016 (X = P, Y =S_p)

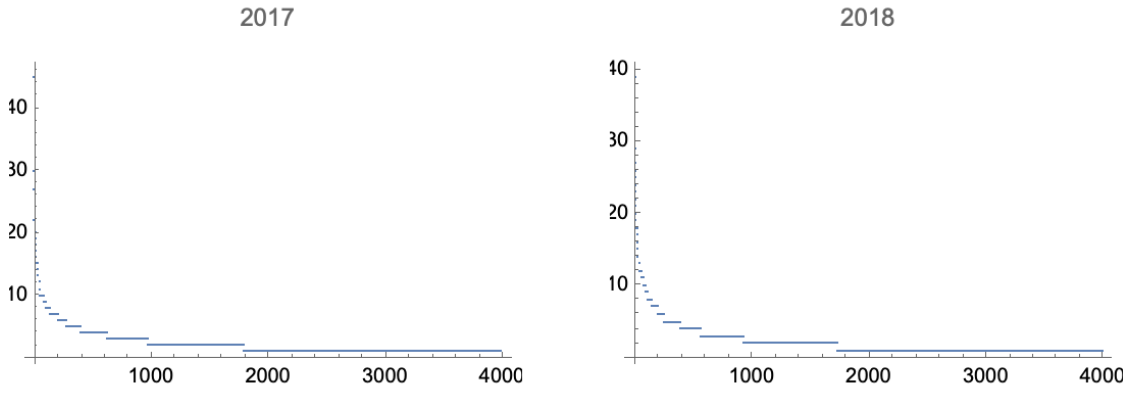

Figure 6: X-Y plot for 2017, 2018 (X = P, Y =S_p)

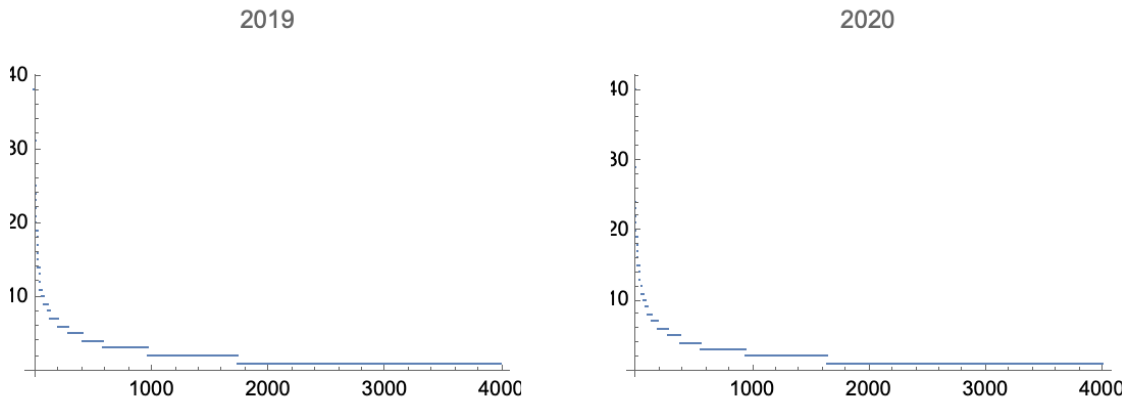

Figure 7: X-Y plot for 2019, 2020 (X = P, Y =Sp)

Stochastic model analysis

1.The result of executing the Borel-Tanner distribution test for each year is shown below.

Year	Parameter Alpha	Parameter N	Pearson-ChiSquare	TestStatistic	PValue	Conclusion
2011	0.57	1	0.01	21.96	0.01	Reject
2012	0.57	1	0.63	7.05	0.63	Do not reject
2013	0.56	1	0.91	3.97	0.91	Do not reject
2014	0.55	1	0.0	23.73	0.0	Reject
2015	0.54	1	0.47	8.7	0.47	Do not reject
2016	0.53	1	0.25	10.29	0.25	Do not reject
2017	0.52	1	0.55	6.84	0.55	Do not reject
2018	0.52	1	0.03	17.02	0.03	Reject
2019	0.53	1	0.19	11.3	0.19	Do not reject
2020	0.53	1	0.13	12.47	0.13	Do not reject

Table 3: Summary the Borel-Tanner distribution tests with conclusions for accepting or rejecting the primary hypothesis.

In 7 out of 10 trials, the null hypothesis can not be accepted at the 5% interval on the basis of the Person Chi Square test. In the remaining 3 trials, the null hypothesis must be accepted.

No other discrete event probability distribution provided any conclusive results.

The selection of "High Needs" locations

From the empirical data and from the generated polynomial model plots it is apparent that a small proportion of physical locations **P** are attributable to a very large service volumes **S** .

I define "high needs" locations as locations $P_{s>5,i}$ which are associated with more than 5 confirmed service visits per month to that location.

Accordingly, the result (count) of selection of P, where $S_p > 5$ are shown below.

$\sum_{i=1}^{10} P_{s>5,i} \; / \; \sum_{i=1}^{10} P_{si} = 3{,}239/20{,}951 = 0.15 = 15\%,$

$\sum_{i=1}^{10} S_{s|p>5,i|} / \sum_{i=1}^{10} S_i = 24319/94333 = 0.257799 =\tilde{} \; 25\%$

Of those locations, $P_{s>5,i}$ I also found that there are 1,273 (39%) locations that are common across i, the entire time period studied. I designate those locations as $\mathbf{P}_{c \mid s > 5}$.

The geospatial histogram of the same high needs locations, $\mathbf{P}_{c \mid s > 5,}$ demonstrates that they are distributed like this, below:

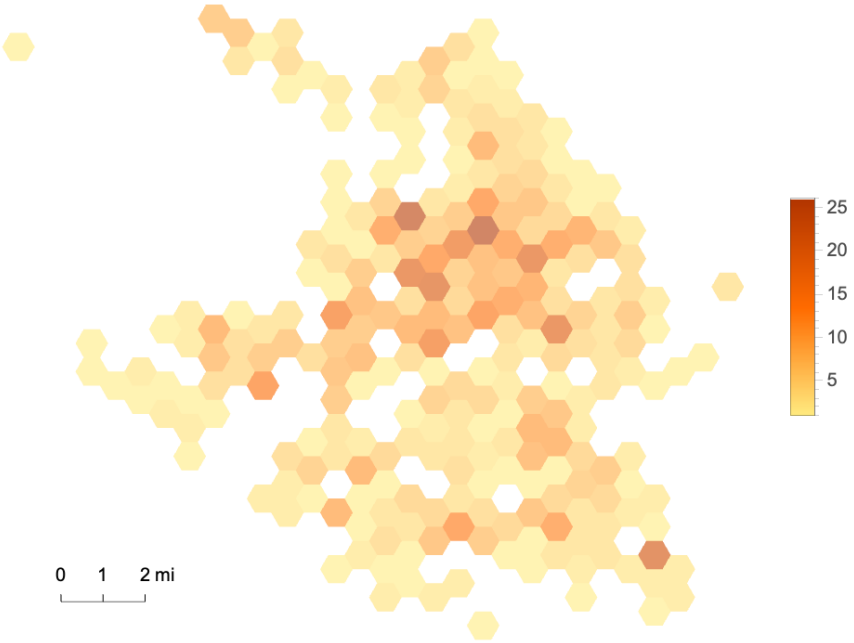

Figure 8: Mapping of high needs physical sites in the City of San Jose (Base Map Removed)

Limitations

There are significant limitations with the source data used in this study. Therefore, the results must be seen with caution and as such, there is no assurance that the results of this preliminary study are appropriate to inform operational decision making or policy.

This study lacks a normative basis against which the source data should be interpreted. For example, I could not find any copy of the metadata catalog issued by the The City of San Jose, nor the SJPD in a manner that conformed with government metadata standards. Therefore, interpretted the data as a 'best guess' from the table field names and from the exploratory data analysis.

There is no assurance provided by the City of San Jose in relation to the quality of the data provided. Therefore this study cannot make any assurance on the quality of the source data.

There are patterns of data that were unexpected. At this stage, I do not have enough specific domain knowledge to explain the unexpected pattern. For example,on the week of May 13, 2021, the volume of recorded police calls increase dramatically. One can only guess that policing intensity may have increased due to more operational budget being allocated in that time period. This has influenced my decision to segment time-ordered data from May of every year and to reject the final year's data.

Address geocoding errors were found due to (a) the use of Common Name in place of Street Names, where Street Names were not provided in the call records. (b) Some Street Names were not standardised. (c) Technical error in the dataset extracted from the Source System (CAD). This means that the study underestimates the number of sites observed. The level of underestimation has not been quantified.

Conclusions

The large observation sample size provided this study the opportunity to develop robust stochastic models with high confidence. The conclusive results from Borel-Tanner tests in 7 out of 10 trials can safely be interpreted as reasonable evidence in support of the primary hypothesis. The limitations of the source data are likely to be a contributing factor for inconclusive results in the other 3 trials.

Literature suggests that Borel-Tanner distributions are commonly found in queueing theory and traffic flow analysis. I speculate that further research into this domain may be able to better determine if call center (and its related call-queuing parameters) operations may be a factor in why such a distribution is observed.

The high proportion of common sites that are highly persistent over time ($\mathbf{P}_{c \mid s > 5,} =\tilde{} 35\%$) relative to all high needs sites ($P_{s> 5}$), is a reasonable explanation of why they appear clustered spatially. I suggest, however, that this effect would need to be better understood in context to other social, economic or physical determinants. Yet, the emergence of the observed clustering effect is likely to assist virtual street audit projects in urban planning decisions.

Conflict of Interest

There are no conflicts of interest that are declared by the author.

References

n.d.

Badland, H.M., Opit, S., Witten, K., Kearns, R.A., Mavoa, S., 2010. Can virtual streetscape audits reliably replace physical streetscape audits?. J Urban Health 87, 1007–16.