# Comparing spatial counts characterizing human mobility using ratio maps

JI MENG LOH*

Given two spatial data sets collected over the same geographical region, a common question of interest is the differences between these two data sets. The data may have been collected at two time points and the differences represent how the data has changed over the intervening time period. Or they may be data on two sub-populations and the differences then describe where and how these two sub-populations vary relative to each other. We consider the problem of objectively highlighting the differences between two spatial data sets, and showing these differences on a map. Specifically, we assume that the data sets are collected over a lattice, and model the values of one set as a function of those of the other data set, taking into account the inherent spatial structure. The model is a hierarchical model with spatial errors whose specification controls the amount of smoothing applied to the data. Additional covariates may be added to the model to explore more complex relationships between the two data sets. We apply the method to study the number of commuters to a town as obtained from cellular Call Detail Records, comparing them to Census data, and also model the commuter numbers with underlying population and distance away from the town. We also study differences in the number of visitors to the town during a St. Patrick's Day parade with visitor numbers on other days.

KEYWORDS AND PHRASES: Bayesian hierarchical model, Differences between maps, Human mobility, Spatial data.

## 1. INTRODUCTION

There is a lot of interest in characterizing human mobility patterns [9, 20]. For example, town planners are interested in traffic patterns and the relative use of roads leading into and out of their towns. Knowledge of the commuting patterns of workers in a city can help urban planners improve transit lines, allocate parking spaces and so on. Understanding where weekend visitors to a town come from can aid planners in marketing its downtown businesses.

Call detail records (CDRs) have recently become an important source of data for the study of human mobility

*Note: much of this work was performed while the author was at AT&T Labs-Research, Florham Park, New Jersey.

[2, 9, 14, 19]. CDRs are records of mobile calls and text messages collected by telecommunications companies for billing purposes. These records contain information about times of calls and the cell towers handling these calls. With the phenomenal rise in the use of mobile phones, this data can provide rich information, on the aggregate level, about the patterns of human mobility. These data are already routinely collected and the cost of preparing the data for further analysis is marginal. Since CDR data is collected constantly as calls are made, there is the potential for studies on the effect of policy changes over short periods of time that would be infeasible or costly to conduct. An example is the effect on traffic patterns of a change in traffic signals on a few intersections or the flow of people into a town for a big event such as a St. Patrick's Day parade.

Since cell towers typically have a coverage area of about 1 square mile [17], it is on a coarse enough scale to leave some measure of uncertainty about the precise location of the phones while providing ample information for studying the aggregate movements of people.

Studies on human mobility often involve counts of people on some form of lattice, e.g. zip-codes. Whether from CDRs or otherwise, the data yield a geographical distribution of people counts based on some pre-specified criteria. With such geographical distributions, what is often of interest is the *comparison* of two or more such distributions. Such comparisons allow investigators to understand how movements of people differ or change under different conditions.

For example, the Census data as contained in the "Journey to Work" portion of the Census Transportation Planning Package (CTPP) provides information about worker residences by zip-code. This data can be used to obtain the *laborshed* of a town, i.e. the geographical distribution of the residence of people who work in that town. A laborshed can provide insight into commuting and traffic patterns around a town and inform policy planning. Figure 1(a) shows a map of the Morristown, New Jersey laborshed using data from the 2000 CTPP. Not surprisingly, most Morristown workers live close to Morristown. However, some workers live further away, e.g. in New York City, and especially in the Newark area between Morristown and New York City.

In [4], CDRs collected for calls (and text messages) made in and around Morristown, New Jersey were used to obtain a similar *laborshed* for the town. Figure 1(b) shows a map of the Morristown laborshed based on the CDR data. Examining both maps in Figure 1 shows that they contain
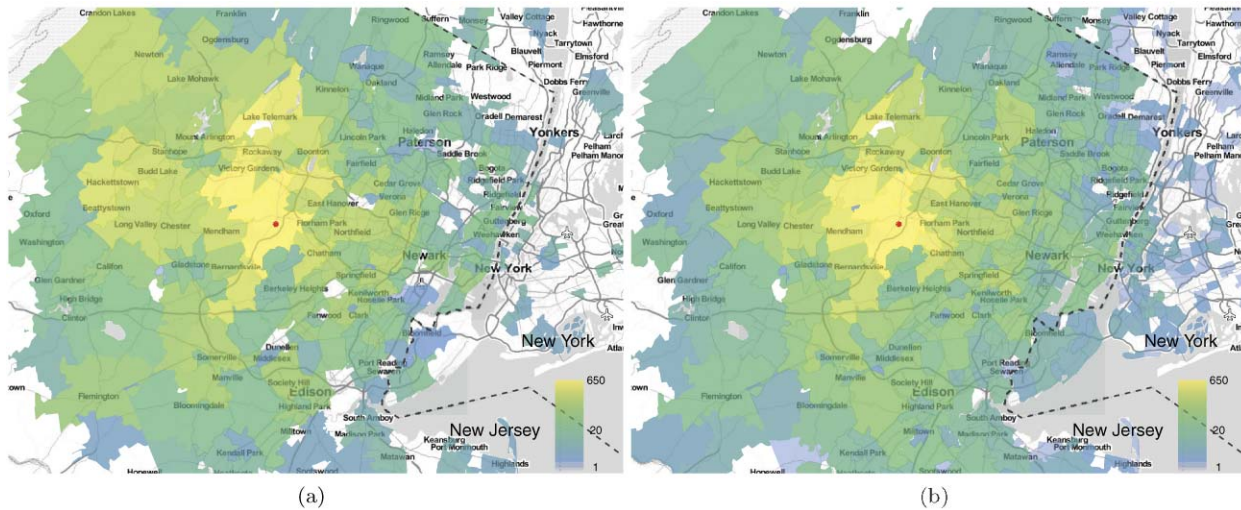
*Figure 1. Maps of the laborshed for Morristown, New Jersey, indicated by the red dot, obtained from the 2000 Census data (left) and derived from Call Detail Records (right). (Color figure online)*

the same main features: a large bump at Morristown, and somewhat higher numbers East of Morristown than to the West.

However, we expect differences between these maps. Both data sets are samples from a population. The underlying populations are also different. With CDRs, the population is of those with cellphones from the company where the data was obtained from, and this population may be biased relative to the population of workers. Furthermore, there is about a 10-year gap between the collection of these two data sets. Comparing the two maps, we note that there appears to be (a) higher CDR numbers in New York City, (b) higher Census numbers further North-West of Morristown, and (c) a few other minor differences, e.g. midway between Manhattan and Morristown.

While an investigator can examine two such maps and try to identify differences between them, this is subjective and different people looking at the same two maps may not identify the same differences. The focus of this paper is thus on objectively and succinctly comparing two spatial distributions, each consisting of counts at a sub-regional (say, zip-code) level, so as to highlight the differences between maps without requiring the reader to subjectively extract this information.

We do this in the context of comparing the Morristown laborsheds obtained from Census data and from CDR data, and of comparing Saturday visitors to Morristown for a St. Patrick's Day parade (a *paradeshed*) to regular Saturday visitors. The latter was studied in [3] and is an example of how CDRs can provide timely information that is not available from Census data and that can be obtained only via a survey study that may be expensive or time-consuming to plan and conduct. Such information about the paradeshed may inform town planners about providing additional transport

services to and from particular neighboring areas in preparation for the parade.

While the focus of the application in this paper is on human mobility, this method of comparing maps can be applied more broadly. For example, in telecommunications, metrics such as the rate of dropped calls, the amount of data uploaded etc. are collected at cell-towers to monitor network performance. The same method described in this work can be used to compare these metrics collected at two time periods to examine, e.g. the differences in performance levels on weekdays and weekends, or to assess the results of improvements made to the network equipment.

We consider two methods for comparing two spatial datasets collected on a lattice. The first method is an ad-hoc method where we aggregate data from nearest neighbors so that the denominator is above some threshold $T$. This helps to control excessive variability due to small denominators. We also construct a quantity as a criterion for selecting $T$. In our second method, we explicitly model the two data sets under comparison, linking one data set with the other using a hierarchical Bayesian model with spatial and independent errors, where the spatial errors have a conditional autoregressive (CAR) specification [6]. The model is fit using a Bayesian approximation method called INLA [15]. The systematic portion of the model shows the main relationships between the two data sets, while the spatial errors show deviations from this systematic relationship. In the comparison of two maps, the spatial errors highlight differences between the two geographical distributions. The modeling approach allows us to include other explanatory covariates that can help in further understanding the relationships between the two spatial sets.

These methods are described in greater detail in Section 4 and used to examine the differences between the CDR and

Census commuter numbers. In Section 5 we use the Bayesian model to incorporate covariates into the modeling, investigating the relation of CDR commuter numbers to the underlying population and the distance from Morristown. We also look at the differences between the St. Patrick's Day *paradeshed* and the geographical distribution of visitors to Morristown on other Saturdays.

## 2. DATA COLLECTION & PRIVACY

Morristown is located in the greater New York City metropolitan area and has about 20,000 residents. It is a regional center of commerce and shopping, with a vibrant night-life, and draws workers from the nearby suburbs and even New York City. The data used in this study comes from Call Detail Records, CDRs for short, collected for calls and text messages made in and around Morristown. More details about the data set can be found in [4].

Several features of the data used in this study ensure individuals' privacy. First, CDRs provide only sparse information about a phone. A record is created only when the phone is used for a call or SMS message. No information about the phone is recorded in the CDRs when it is not in use. Also, the location of a phone is available only approximately, based on the coverage of the antennas involved with the call, and not with GPS. Since, as mentioned earlier, an antenna typically covers an area of about one square mile, this limits the spatial resolution available on individual phones.

Besides the inherent features of the data, we also took active steps to protect privacy. Specifically, all CDRs are anonymized by a third party not involved in this work. The phone number is not used. Rather, a separate unique identifier is generated and used in this analysis.

In addition, we are careful to use only the minimal information needed for the study. Our data records consist only of the anonymous identifier, date and time of a voice call or SMS message, elapsed time of a call, and the cell antennas involved in the event. Our data does not include demographic information for the cellular subscriber nor any information about the other party in the communication.

Finally, all of our results are presented in aggregate form, specifically at the zip-code level. No individual phone was singled out for analysis.

## 3. RELATED WORK

Work done in disease mapping is relevant to our work. There is much research in this area and this brief overview cannot be exhaustive. The reader can refer to [11] and [10] for more comprehensive overviews of common methods. In disease mapping, the interest is often to display the relative risk, also called the standardized mortality rate (SMR). This is given by the ratio of the observed to the expected number of disease cases. Similar to our comparison of two spatial data sets, the SMR suffers from increased variability due to small denominators.

[7] introduced an empirical Bayes method to shrink relative risk estimates towards the global estimate. [13] extended this to allow for spatial autocorrelation. Hierarchical models have been applied to epidemiological data to estimate relative risk. [12] compared a spatial filtering method and a binomial hierarchical model for mapping the spatial relative risk of neural tube defects for several provinces in China.

[5] considered using kriging to obtain a continuous spatial risk function from regional count data. [8] proposed a method to better estimate the relative risk of diseases using an adaptive bandwidth method for density estimation that is specially catered for estimating relative risks (i.e. the ratio of risks). Briefly, they defined the log-transformed risk as $\rho = \log(f/g)$ where $f$ and $g$ are densities, each estimated by a kernel density estimator. So an estimator for $f$ is given by

$$\hat{f}(z) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_f(x_i)^2} K\left(\frac{z - x_i}{h_f(x_i)}\right),$$

where $K$ is a kernel and $h_f$ the bandwidth. Thus $\hat{\rho}(z) = \log[\hat{f}(z)/\hat{g}(z)]$. Their adaptive method involves letting $h_f$ and $h_g$ vary, so that e.g.

$$h_f(x) = \frac{h_0}{f(x)^{1/2}\gamma},$$

with $h_0$ a global bandwidth and $\gamma$ the geometric mean of $f$. Since $f$ and $g$ are not known, $h_f$ and $h_g$ are estimated using pilot estimates of $f$ and $g$ obtained with fixed bandwidths. [8] also provides the asymptotic bias and variance of the resulting risk estimator.

The adaptive bandwidth method of [8] is available in an R package *sparr*. We applied it to our data sets but the method did not seem to work well, with extreme values around the borders of our study region. This suggests that the reason could be due to inadequate accounting of the complicated boundaries of the study region, though one would expect the edge effects in the numerator and denominator to cancel out. Other adaptive methods, e.g. based on Voronoi tessellations in the *spatstat* R package [1], are available, but they are not explored here.

Another possible method is to use spatial interpolation. To do that we can use the *smooth* function in the *spatstat* package. However, this function performs kernel smoothing in the numerator and the denominator and encountered problems when the denominator is small. A possible extension would be to perform adaptive kernel estimation for the interpolation.

In the paper we use a Poisson hierarchical model and fit the model using a Bayesian approximation method called INLA. [16] also considered using INLA for disease mapping. Their focus, however, was on models for spatial-temporal data with emphasis on modeling the time trend, parametrically or non-parametrically.
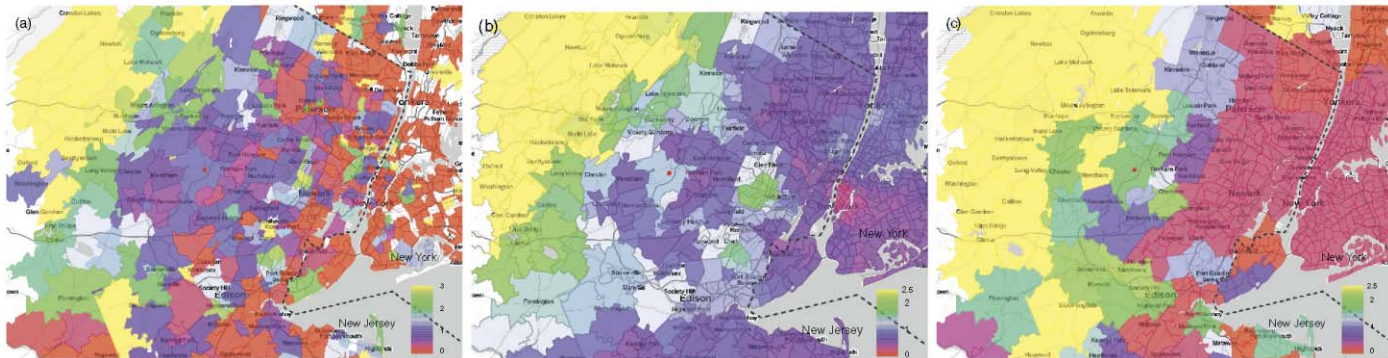
Figure 2. Ratio of Census to CDR counts on the zip-code level of commuters to Morristown, NJ. (a) raw ratios, (b) ratios smoothed by using neighboring zip data to make the denominator at least a threshold level $T = 300$, and (c) ratios obtained from the hierarchical model fit using INLA. (Color figure online)

## 4. METHODS

In this section, we describe two methods for comparing two spatial data sets on a lattice.

### 4.1 Thresholds for denominators

One way to compare two spatial data sets is to simply take the ratio of the data values at the zip-code level and plot the ratios on a map. We choose a color scheme that has light blue in the middle of the scale, and green-yellow and purple-red on the opposite ends of the scale. Doing this allows an easy way to visually show large and small values of the ratios.

Applying this to the CDR and Census data results in the leftmost map of Figure 2. While this figure shows accurately the differences between the two data sets, there are some extreme values at individual zip-codes, due to small denominators. To deal with this, we perform some smoothing. Suppose for zip-code $z$, the data values are $u_z$ and $v_z$. We set a minimum threshold $T$ for the denominator. For each $z$, if $v_z < T$, we consider the nearest zip-code $z'$. If $v_z + v_{z'} \geq T$, then we set the ratio to be $r_z = (u_z + u_{z'})/(v_z + v_{z'})$. Otherwise we consider the next nearest zip-code and so on, until the sum of the $v$ values of zip-code $z$ and the neighboring zip-codes is at least $T$. The resulting values are denoted $u_z^T$ and $v_z^T$ and the ratio that is plotted at zip-code $z$ is given by $u_z^T/v_z^T$.

In order to find an appropriate $T$, we can use a criterion function that consists of two terms, one increasing in value and the other decreasing in value as $T$ increases:

$$(1) \quad C(T) = \sum_z \left( \frac{u_z^T}{v_z^T} - \frac{u_z}{v_z} \right)^2 + \sum_z \left( \frac{u_z^T}{v_z^T} - \frac{\sum_z u_z}{\sum_z v_z} \right)^2.$$

The expression in the brackets of the first term on the right-hand side of (1) represents the bias between the smoothed ratio $u_z^T/v_z^T$ and the raw ratio $u_z/v_z$. The quantity $\sum_z u_z / \sum_z v_z$ represents the overall ratio between the

two datasets when there is complete smoothing and no variation between zip-codes $z$. The value $T_{\min}$ that minimizes $C(T)$ balances between not deviating too much from the overall ratio while staying true to the actual zip-code values. With the Census and CDR commuter data, $T_{\min} = 302$. The resulting map is shown in Figure 2(b) for $T_{\min} = 300$.

We find that the main features showing the differences between the two laborsheds are still present but with extreme values present in the map on the left smoothed out. There is somewhat more smoothing to the West than to the East in New York City. The zip-codes in New York City have relatively large populations but have small areas. Hence the neighboring zip-codes included to make the denominator larger than $T$ will tend to cover a smaller spatial area in New York City than further inland in New Jersey.

### 4.2 Bayesian smoothing

Instead of the threshold method described above, we can compare two maps using Bayesian smoothing. Specifically, for data $U_z$ and $V_z$ we have

$$U_z \sim \text{Poisson}(\mu_z)$$
$$\log(\mu_z) = \alpha + \log V_z + b_z + e_z,$$

where $b_z$ and $e_z$ are spatial and independent errors respectively. The prior for $e_z$ is set to be independent normal, while for $b_z$ we use a conditional autoregressive (CAR) specification [6], where the neighborhood matrix is defined using a distance threshold of 5 miles, i.e. a zip-code's neighbors are all zip-codes within a 5 mile radius of that zip-code. We note that Manhattan is about 2.5 miles across at its widest, hence 5 miles spans a distance slightly less than 1/10 of the map in Figure 2.

We can fit the model in two ways, using the Open-BUGS software for a full Bayesian fit or the approximation Bayesian method INLA developed by [15]. We chose the INLA method as it allows for very quick fitting of the model to the data. The map on the right of Figure 2(c) shows the

fitted values of $\mu_z/V_z$ obtained from the INLA method. We find that the map obtained from Bayesian smoothing, like the one obtained with the threshold method of Section 4.1, displays differences between the two maps in Figure 1 in a smooth way that avoids sharp jumps in values.

## 5. BAYESIAN MODELING

The advantage of the Bayesian smoothing method of Section 4.2 is that the Bayesian model can be used to explore dependencies of the data on various covariates. We show here modeling done to explore how the numbers of commuters to Morristown relative to the underlying zip-code level population varies with distance from Morristown.

The model is:

$$
\begin{aligned}
(2) \qquad U_z &\sim \text{Poisson}(\mu_z) \\
\log(\mu_z) &= \alpha + \log p_z + \beta d_z + b_z + e_z,
\end{aligned}
$$

where $U_z$ is the number of commuters to Morristown from zip-code $z$ inferred from the CDR data, $p_z$ the zip-code population obtained from the 2000 Census data, and $d_z$ the distance from the zip-code centroid to Morristown. Maps in Figure 3 show the results of fitting this model. The Deviance Information Criterion (DIC), a measure of model fit introduced by [18], dropped from 3,402 for the model without $d_z$ to 3,385 for (2).

Figure 3(a) shows the fitted mean values $\mu_z$, with yellow, green and light blue indicating decreasing numbers. We find that most workers in Morristown reside in the neighborhood of Morristown, with a higher tendency to be just north and south of Morristown than to its east or west. Figure 3(b) shows the count per 6,000 people in the zip-code. This accounts for the underlying population density. Although we find a similar concentration at Morristown, this second map shows a clearer circular shape, with counts decaying with distance. Comparing Figure 3(a) with 3(b) suggests that the areas just north and south of Morristown have a higher population which contributes to the higher counts in the first map. The estimate of $\beta$ is $-.159$, which gives the expected rate of decay with distance of the proportion of a zip-code's population that work in Morristown: with each mile away from Morristown, the proportion of a zip-code's population that work in Morristown drops by about 15%.

Figure 3(c) is a map of the spatial errors where the yellow-green scale indicates positive errors while the red-purple scale shows negative errors. It shows locations where there are more (or fewer) commuters to Morristown after accounting for distance and population. There is a distinct separation between the areas east and west of Morristown, suggesting that there is a difference in the numbers of commuters to Morristown (relative to underlying population) between the areas east and west of Morristown, with a higher proportion of the population west of Morristown working in Morristown. Morristown is one of the larger towns in the area and thus attracts a significant number of workers from its

*Table 1. Coefficient estimates, standard errors and 95% intervals for models (2) and (3) using commuter counts based on CDR data (top two lines and bottom three lines respectively)*

|  | estimate (error) | 95% interval |
|---|---|---|
| intercept | $-3.68$ (.258) | $(-4.18, -3.16)$ |
| $\beta$ | $-.159$ (.009) | $(-.177, -.142)$ |
| intercept | $-3.765$ (.252) | $(-4.26, -3.26)$ |
| $\beta_{\text{east}}$ | $-.163$ (.009) | $(-.188, -.146)$ |
| $\beta_{\text{west}}$ | $-.115$ (.010) | $(-.136, -.095)$ |

neighboring vicinity. However, working people living east of Morristown also live closer to New York City, a major source of jobs and is probably the reason for a smaller proportion of these people working in Morristown.

Figure 3(d) shows the independent errors obtained from model (2) and shows much less spatial structure than Figure 3(c) as expected.

We can fit a model allowing for different rates of decay with distance east and west of Morristown:

$$
\begin{aligned}
(3) \qquad U_z &\sim \text{Poisson}(\mu_z) \\
\log(\mu_z) &= \alpha + \log p_z + \beta_{\text{west}} I_{\text{west},z} d_z \\
&\quad + \beta_{\text{east}} I_{\text{east},z} d_z + b_z + e_z,
\end{aligned}
$$

$I_{\text{west},z}$ and $I_{\text{east},z}$ being indicator variables, with value 1 if zip-code $z$ is west or east of Morristown respectively. We find a slight drop in the rate of decay for zip-codes west of Morristown, with rate $-.115$ compared to $-.163$, in agreement with our observation from the maps. The DIC for this model is 3,363. Note from Table 1 that the 95% intervals for the rates east and west of Morristown do not overlap.

Figures 3(e) and (f) show respectively the spatial and independent errors obtained from fitting (3). The spatial errors still show mostly negative values in the New York area, but generally this is a bit less pronounced than the spatial error map from fitting (2). Table 1 shows the coefficient estimates and standard errors for models (2) and (3).

### 5.1 Modeling St. Patrick's Day parade-goers

We perform a similar analysis to study the geographical distribution of attendees of a St. Patrick's Day parade in Morristown, comparing it to visitors to Morristown on other Saturdays. To do this, we focus on calls and messages made during the parade, from 11am to 3pm on Saturday Mar 12, 2011, that were handled by cell-towers covering the parade route, and on the same time period on other Saturdays in the dates covered by our data. If $P_z$ and $S_z$ represent the numbers of visitors to Morristown from zip-code $z$ on Mar 12, 10am–3pm and on other Saturdays respectively, the two models we use are given by

$$
\begin{aligned}
(4) \qquad P_z &\sim \text{Poisson}(\mu_z) \\
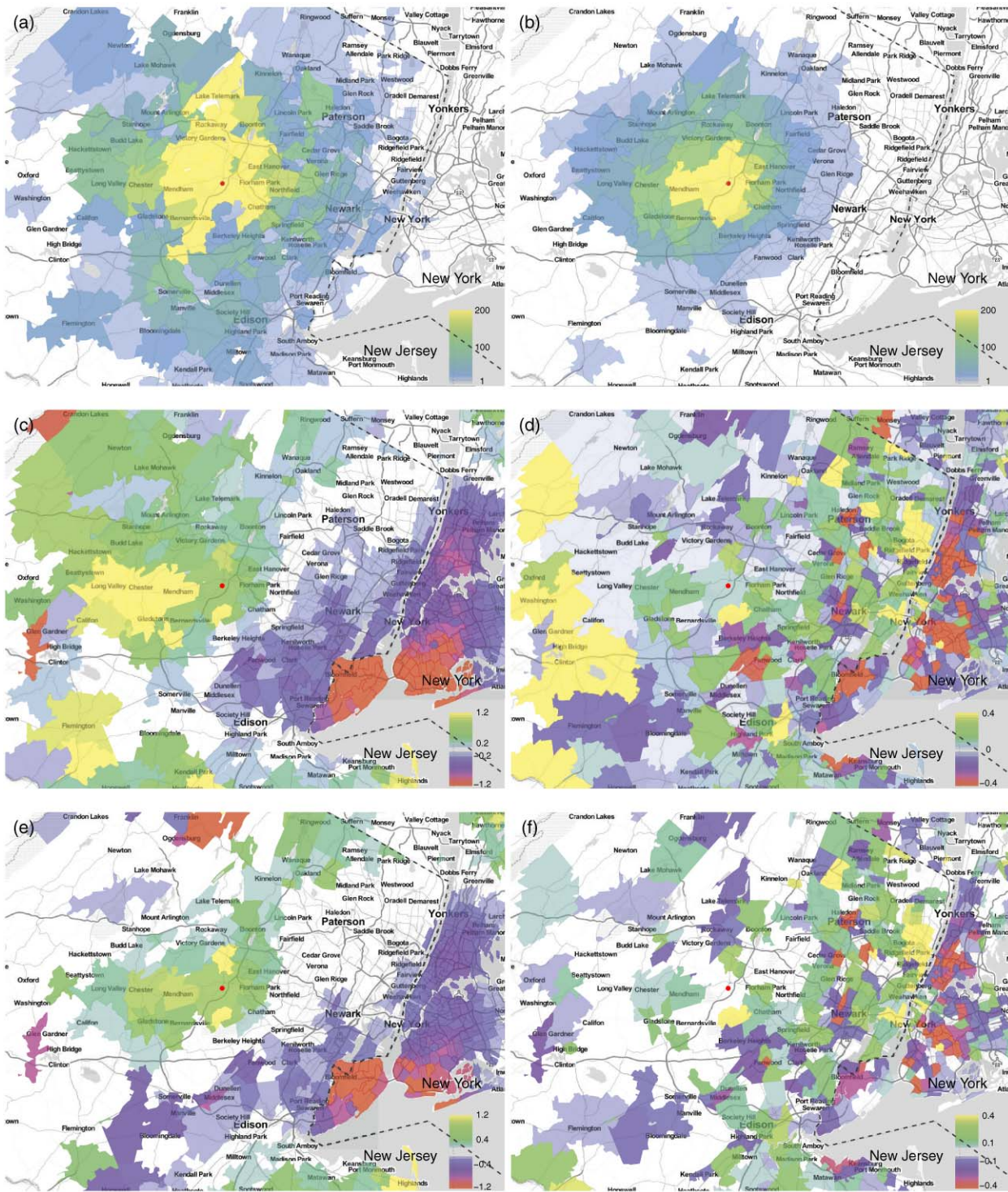\log(\mu_z) &= \alpha + \log(S_z) + b_z + e_z
\end{aligned}
$$

Figure 3. Maps at the zip-code level for the CDR commuter data. From top to bottom, left to right, these are maps of (a) smoothed counts, (b) smoothed counts per 6,000 people, (c) spatial errors with the distance covariate, (d) independent errors with the distance covariate, (e) spatial errors with East/West distance covariate and (f) independent errors with East/West covariate. (Color figure online)
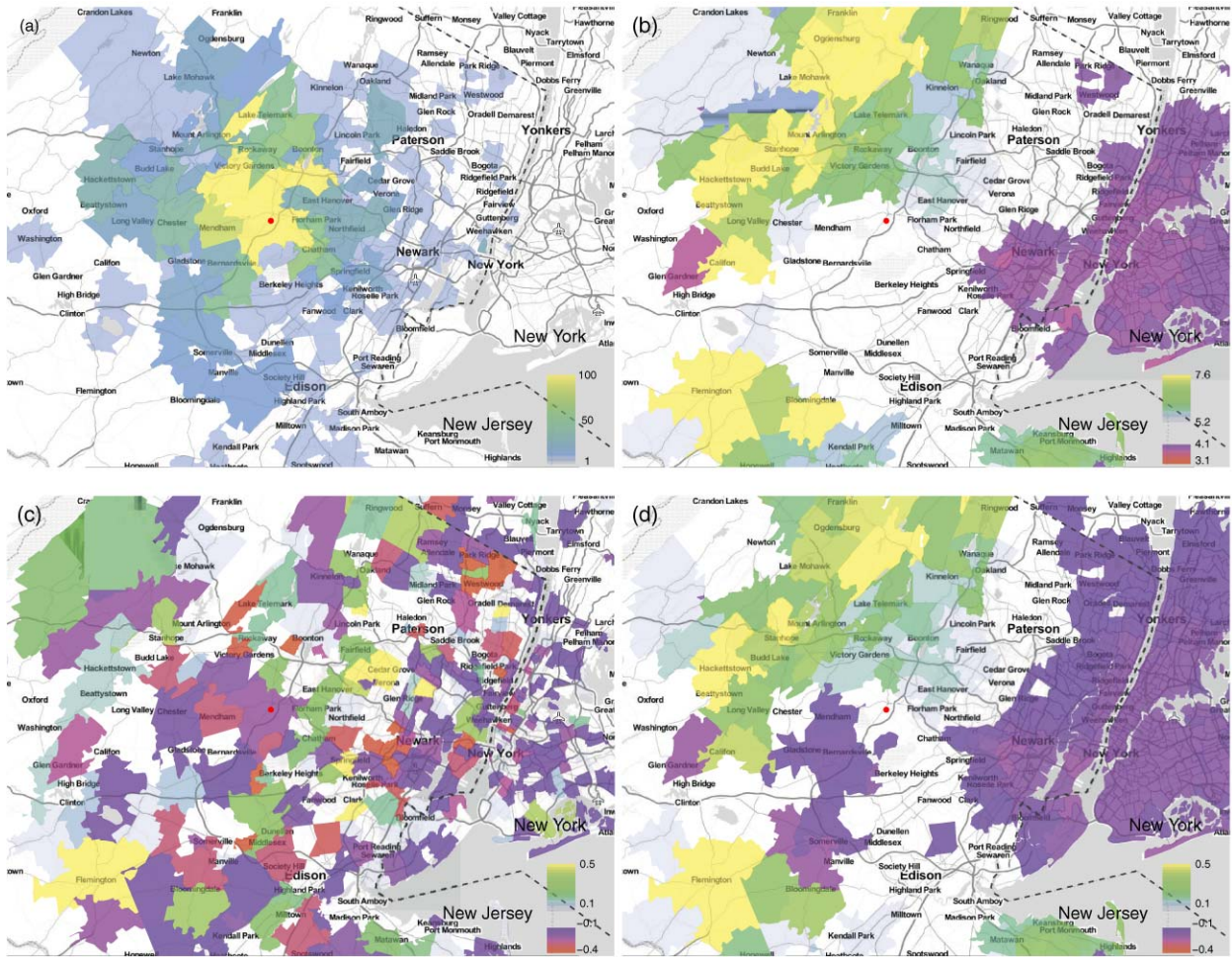
Figure 4. Maps at the zip-code level comparing the geographical distribution for attendees of the St. Patrick's Day parade to that of regular Saturday visitors to Morristown, NJ. From top to bottom, left to right, these are maps of (a) smoothed counts of parade-goers, (b) ratio of parade goers to regular visitors, (c) independent errors, (d) spatial errors from the model with a distance covariate. (Color figure online)

and

$$(5) \qquad \log(\mu_z) = \alpha + \log(S_z) + \beta I_{\text{west},z} + b_z + e_z.$$

The maps are shown in Figures 4(a)–(d). Figure 4(a) shows the smoothed counts of parade-goers. As expected, most of them come from Morristown and its vicinity. Figure 4(b) compares the counts by zip-code of parade-goers compared with average counts of visitors to Morristown during the same time period on other Saturdays. We find an increased number of parade-goers to the north and west of Morristown and lower than usual visitors on St. Patrick's Day from the New York City area. Figure 4(c) shows the independent errors from model (4), showing little spatial structure. Finally, the spatial errors obtained from model (5) fit to the data are shown in Figure 4(d). We find that there is still a fair amount of spatial structure, suggesting that the additional covariate did not capture well the residual spatial structure from model (4). Indeed, the 95% predictive interval obtained

from INLA contained 0. The DIC of the two models are also very similar.

## 6. CONCLUSION

It is common when dealing with spatial data sets to compare the distribution of these data sets in order to identify differences or changes between them. We proposed a method to objectively highlight these differences. The method involves modeling one data set as a function of the other together with spatial and independent errors. The spatial errors have a conditional autoregressive specification, and the amount of smoothing can be controlled by means of the neighborhood structure. Computationally efficient fitting of the model is achieved using the INLA Bayesian approximation method. The model also allows for additional covariates to be included so that more complex relationships between the data sets can be explored.

In our analysis, we found that Morristown commuter counts decreased with distance from Morristown, as expected. In particular, we found that the decay in counts with distance was greater east of Morristown (nearer New York City) than west of it. With the data on visitors to Morristown on Saturday mornings, we find that areas to the north-west of Morristown had more parade attendees than expected based on visitor counts from other Saturdays.

Measures of uncertainty can be as important as maps of the smoothed ratios. Although we do not reproduce them here, maps showing the standard errors can be drawn using the output of the model fit.

There is some subjectivity in terms of the choice of colors and the range of values over which the colors represent. However, this is true when drawing any map.

## REFERENCES

[1] BADDELEY, A. and TURNER, R. (2005). Spatstat an R package for analyzing spatial point patterns. *Journal of Statistical Software* **12** 1–42.

[2] BARABASI, A.-L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature* **435** 207–211.

[3] BECKER, R. A., CÁCERES, R., HANSON, K., ISAACMAN, S., LOH, J. M., MARTONOSI, M., ROWLAND, J., URBANEK, S., VARSHAVSKY, A., and VOLINSKY, C. (2011a). Human mobility characterization from cellular network data. Submitted to *CACM*.

[4] BECKER, R. A., CÁCERES, R., HANSON, K., LOH, J. M., URBANEK, S., VARSHAVSKY, A., and VOLINSKY, C. (2011b). A tale of one city: Using cellular network data for urban planning. *IEEE Pervasive Computing* **10** 18–26.

[5] BERKE, O. (2004). Exploratory disease mapping: Kriging the spatial risk function from regional count data. *International Journal of Health Geographics* **37**.

[6] BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society Series B* **36** 192–236. MR0373208

[7] CLAYTON, D. and KALDOR, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* **43** 671–681.

[8] DAVIES, T. M. and HAZELTON, M. L. (2010). Adaptive kernel estimation of spatial relative risk. *Statistics in Medicine* **29** 2423–2437. MR2759957

[9] GONZÁLEZ, M. C., HIDALGO, C. A., and BARABÁSI, A.-L. (2008). Understanding individual human mobility patterns. *Nature* **453** 779–782.

[10] LAI, P.-C., SO, F.-M., and CHAN, K.-W. (2009). *Spatial Epidemiological Approaches in Disease Mapping and Analysis.* CRC Press, New York.

[11] LAWSON, A. B. and WILLIAMS, F. L. R. (2001). *An Introductory Guide to Disease Mapping.* John Wiley, New York.

[12] LIAO, Y.-L., WANG, J.-F., WU, J.-L., WANG, J.-J., and ZHENG, X.-Y. (2011). A comparison of methods for spatial relative risk mapping of human neural tube defects. *Stochastic Environmental Research and Risk Assessment* **25** 99–106.

[13] MARSHALL, R. (1991). Mapping disease and mortality rates using empirical Bayes estimators. *Applied Statistics* **40** 283–294.

[14] PHITHAKKITNUKOON, S., SMOREDA, Z., and OLIVIER, P. (2012). Socio-geography of human mobility: A study using longitudinal mobile phone data. *PLoS ONE* **7** e39253.

[15] RUE, H., MARTINO, S., and CHOPIN, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B* **71**(2) 319–392. MR2649602

[16] SCHRÖDLE, B. and HELD, L. (2011). Spatio-temporal disease mapping using INLA. *Environmetrics* **22** 725–734. MR2843139

[17] SONG, C., QU, Z., BLUMM, N., and BARABÁSI, A.-L. (2010). Limits of predictability in human mobility. *Science* **327** 1018–1021. MR2643139

[18] SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P., and VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B* **64** 583–616. MR1979380

[19] WESOLOWSKI, A., EAGLE, N., NOOR, A. M., SNOW, R. W., and BUCKEE, C. O. (2013). The impact of biases in mobile phone ownership on estimates of human mobility. *Journal of the Royal Society Interface* **10**.

[20] YUAN, J., ZHENG, Y., and XIE, X. (2012). Discovering regions of different functions in a city using human mobility and POIs. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'12, pp. 186–194, ACM, New York, NY, USA.

Ji Meng Loh
Dept of Mathematical Sciences
New Jersey Institute of Technology
University Heights, Newark, NJ 07102
USA
E-mail address: loh@njit.edu