CitySenseTM: multiscale space time clustering of GPS points and trajectories

Markus Loecher¹ and Tony Jebara^{1,2} ¹Sense Networks, 110 Greene Str., NY, NY 10012 ²Columbia University, 1214 Amsterdam Ave, NY, NY 10027

Abstract

A generalized scan statistic is provided for periodic geographic surveillance of various measures of city-wide activity. The resulting algorithms scan for both elliptical and rectangular clusters and can be optionally adjusted for covariates such as time-of-day, day-of-week or season. An empirical Bayes procedure is used to account for parameter uncertainties in the framework. Additionally, scanning is performed to recover second order space-time clusters corresponding to transition dynamics of agents moving within a city. The method simultaneously evaluates the existence and significance of primary source and secondary destination clusters. It enables the user to address the question: "given a location in time and space, which – if any- are the most likely circular/elliptical/rectangular regions that agents usually visit subsequently". Computational efficiency is maintained by using a multi-scale grid that is adapted to the geographic area being scanned.

Our mobile application, **CitySense**, is currently analyzing real-time feeds plus a billion points of GPS and WiFi positioning data from the last few years in San Francisco to provide a map based summary of current hotspots of activity.

Key Words: Scan Statistic, spatiotemporal hot spots, clustering, anomaly detection

1. Introduction

Our method expands upon and applies the space time statistics proposed by Kulldorff to collections of GPs points in major metropolitan areas. It performs periodic geographic surveillance of various, spatially distributed measures of interest by scanning for spatial or spatiotemporal clusters and evaluates their statistical significance. These measures can be discrete events, e.g. Scan statistics are applied in various scientific disciplines to model clusters of events in space and time. These methods provide a principled way of testing whether an observed cluster of events occurs by chance given a probability model. This article expands upon and applies the space time statistics originally proposed by Kulldorff to GPS sample data in multiple cities.

The resulting system, referred to as CitySense, performs periodic geographic surveillance of various, spatially distributed measures of interest by scanning for spatial as well as spatiotemporal clusters and evaluating their statistical significance. These measures can be discrete events (e.g. arrivals or departures of taxi cabs or other vehicles) as well as continuous variables (travel times or demographic attributes). The system scans for circular, elliptical as well as rectangular clusters by combining the methodologies by Kulldorff and Neill/Moore. In an optional step, the method adjusts for covariates such as time-of-day, day-of-week and/or season

via statistical modeling, thus empowering the user to only scan for clusters that deviate from established patterns. Another extension of the system permits it to search for second order space-time clusters in the transition dynamics of agents moving within a city. The method simultaneously evaluates the existence and significance of primary source and secondary destination clusters. This permits the system to predict, given an agent's location in time and space, the likely subsequent locations in time and space the agent will visit. Currently, the CitySense system transmits its output in near real-time on handheld mobile phone devices allowing thousands of users in a given city to examine activity levels and facilitates a variety of social applications.

1.1 CitySense on a mobile device:

Citysense is an application that operates on the Sense Networks Macrosense platform, which analyzes massive amounts of aggregate, anonymous location data in real-time. Macrosense is already being used by business people for things like selecting store locations and understanding retail demand. Citysense uses all this real-time data to enhance nightlife searches. Current platforms supported are Apple iPhone®, Apple iPod® Touch andBlackBerry® devices such as Pearl (models 8100, 8120, 8130) Gamma Ray (8800, 8820, 8830), Curve (8300, 8310, 8320, 8330).

CitySense provides an overall activity index for the entire city of San Francisco and compares it to expected values that are based on seasonally and covariate-adjusted historical data.

"Live overall activity & top hotspots

First of all see if it's a good night to go out. The city is 21% busier than normal for right now? Let's go. But where to? Check out the top hotspots in real-time and head out."

"Show me where the unusually high activity is Even if you're a local, Citysense can give you the live details you need. When the Mission or Soma is busier than normal - you'll know immediately."

"Find out where everyone is going

After dinner, drinks or a great night at a club, do you ever wonder where the afterparty is? Just press the "Locate Me" icon and see the top 5 places people go to from where you are now"



Fig.1: Introduction to the mobile app CitySense, at www.citysense.com

In addition, CitySense can show a city's nine busiest locations using a color-coded scale, with red being the most popular. It also displays a relative "busy" rating to indicate whether users' cities are busier or more calm than usual. And a single click on a specific area opens up a directory of

nearby nightlife activities and establishments from Yelp and Google.

2. Background

CitySense is a system for geospatial hot-spot detection, where hot-spot refers to an anomaly, outbreak or elevated/depleted cluster. While CitySense is focused on real-time social monitoring of a city, it relates to other similar efforts in diverse fields including epidemiology, remote sensing, environmental statistics, biosurveillance, and ecology. In general, monitoring spatially and temporally varying activity of various kinds is an important tool for many technologies and scientific disciplines. The spatial scan statistic (Kulldorff and Nagarwalla, 1995) and its associated public domain software SatScan[™] is widely used for the detection and evaluation of disease clusters. CitySense constitutes both a novel application as well as a significant improvement of the existing state-of-the-art algorithms identifying areas of exceptionally high-dimensional settings; in practice it is most suitable for the detection of purely temporal (one dimensional), purely spatial (one to three dimensional) and space-time (two to four dimensional) clusters.

Currently available spatial scan statistics software employ circular, elliptical or rectangular scanning windows and offer various distributions for the Null hypothesis of no clustering. SatScan[™] in particular uses either a non-homogeneous Poisson-based model, where the number of events in a geographical area is Poisson-distributed, according to a known underlying population at risk; a Bernoulli model, with 0/1 event data such as cases and controls; a space-time permutation model, using only case data; an ordinal model, for ordered categorical data; an exponential model for survival time data with or without censored variables; or a normal model for other types of continuous data.

The following shortcomings are addressed by CitySense:

- Over-dispersed count data: With its fixed mean-variance relation, the Poisson model is rather restrictive. In practice, over-dispersed (variance >> mean) count data are encountered quite frequently leading to a high number of spurious clusters in the standard model.
- (ii) Within-cluster modeling: The original spatial and space-time scan statistic define a very specific (multiplicative) model describing the activity inside a cluster which can be too rigid in many applications.
- (iii) Spatial correlations: There is no possibility to adjust for existing spatial correlations which also leads to inflated significance p-values.
- (iv) Prior knowledge: System parameters such as the expected incidence rates (Poisson/Bernoulli model) or mean/variance (exponential/normal model) are estimated from the data presented to the software. When prior knowledge exists, one should be able to supply those parameters instead.

3. Detailed Description:

CitySense periodically scans for space-time clusters in a set of GPS points collected by agents (taxi cabs, cars, trucks, mobile phones, etc.) and their derivatives, while optionally adjusting for covariates such as weather, holidays, time-of-day, day-of-week and seasonality.

The system involves data either in a spatially continuous representation or using a grid based approach (by aggregating the variables of interest to predefined cells in a spatial grid). Such a grid can be a regular rectangular partition, a Voronoi tiling based upon data density, or correspond to natural geographic entities such as provinces, counties, parishes, census tracts, postal code areas, school districts, households, etc.

CitySense processes data typically as counts which are assumed to follow a negative binomial distribution NB(\mathbf{b}_i , \mathbf{o}_i). More formally, a set of grid cells U= { $\mathbf{s}_1 \dots \mathbf{s}_T$ } are provided (from a rectangular, Voronoi or manual partitioning of the sample space). Subsequently, every grid cell \mathbf{s}_i is associated with a baseline (or expected count) \mathbf{b}_i as well as an over dispersion parameter $\mathbf{o}_i \ge 1$ which parameterizes the excess variance over the Poisson distribution: variance = $\mathbf{b}_i * \mathbf{o}_i$. Assuming that the observed counts \mathbf{c}_i in each cell are independently NB distributed, the goal of CitySense is to find spatial regions S with counts significantly greater than the baseline. In the framework of hypothesis testing, we test the null hypothesis H₀ against the set of alternative hypotheses H_a(S), where:

H₀: $\mathbf{c_i} \sim \text{NB}(\mathbf{b_i}, \mathbf{o_i})$ for all cells.

 $H_a(S)$: $c_i \sim NB(q^*b_i, o_i)$ where q > 1, for all s_i in S and $c_i \sim NB(b_i, o_i)$ outside S. We compute the likelihood ratio $LR(S) = Pr[Data | H_a(S)]/Pr[Data | H_0]$ using the maximum likelihood estimate for the parameter q:

$$LR(S) = \frac{\max_{q>1} \prod_{s_i \in S} \Pr[c_i \sim NB(q \cdot b_i, o_i)]}{\prod_{s_i \in S} \Pr[c_i \sim NB(b_i, o_i)]}$$

As the corresponding likelihood ratio for the Poisson case simplifies substantially:

$$LR(S) = \frac{\max_{q>1} \prod_{s_i \in S} \Pr[c_i \sim Pois(q \cdot b_i)]}{\prod_{s_i \in S} \Pr[c_i \sim Pois(b_i)]} = \frac{\max_{q>1} \prod_{s_i \in S} q^{c_i} e^{-qb_i}}{\prod_{s_i \in S} e^{-b_i}} = \frac{\max_{q>1} q^C e^{-qB}}{e^{-B}} = \max\left[1, \left(\frac{C}{B}\right)^C e^{B-C}\right]$$

(with $C = \sum_{s_i \in S} c_i$ and $B = \sum_{s_i \in S} b_i$) we apply the following variance stabilizing transformation $\tilde{c}_i = F_{Pois}^{-1} [F_{NB}(c_i, b_i, o_i), b_i]$ where F_{Pois} and F_{NB} are the cumulative distribution functions (CDFs) for the negative binomial and Poisson distribution, respectively. This function translates counts from a negative binomial to a Poisson distribution while preserving the mean rate.

If the baseline rates are not known a priori, one defines the population based scan statistic as

$$LR_{pop}(S) = \max\left[1, \left(\frac{C_{in}}{B_{in}}\right)^{C_{in}} \left(\frac{C_{out}}{B_{out}}\right)^{C_{out}} \left(\frac{C_{all}}{B_{all}}\right)^{-C_{all}}\right]$$

where in a natural extension, $[C_{in}, B_{in}]$, $[C_{out}, B_{out}]$ and $[C_{all}, B_{all}]$ denote the total count and baseline inside and outside region S and everywhere, respectively.

Alternatively, if the variables of interest follow a continuous distribution, we compute the likelihood ratio for e.g. the Gaussian or exponential distributions. For the following set of hypothesis

H₀: $\mathbf{c}_{\mathbf{i}} \sim \text{Gaussian}(\boldsymbol{\mu}_{\mathbf{i}}, \boldsymbol{\sigma}_{\mathbf{i}})$ for all cells.

H_a(S): $\mathbf{c}_{i} \sim \text{Gaussian}(\mathbf{q}\boldsymbol{\mu}_{i}, \sigma_{i})$ where q > 1, for all \mathbf{s}_{i} in S and $\mathbf{c}_{i} \sim \text{Gaussian}(\boldsymbol{\mu}_{i}, \sigma_{i})$ outside S. Following Kulldorff and Neill, we derive the analytic expression for LR(S):

$$LR(S) = \frac{\max_{q>1} \prod_{s_i \in S} \Pr[c_i \sim N(q \cdot \mu_i, \sigma_i)]}{\prod_{s_i \in S} \Pr[c_i \sim N(\mu_i, \sigma_i)]} = \frac{\max_{q>1} \prod_{s_i \in S} \exp[-(c_i - q \cdot \mu_i)^2 / 2\sigma_i^2]}{\prod_{s_i \in S} \exp[-(c_i - \mu_i)^2 / 2\sigma_i^2]}$$
$$= \max_{q>1} \exp[(1 - q^2)\tilde{B}/2 + (q - 1)\tilde{C}] = \max\left[1, \exp\left(\frac{C^2}{2B} + \frac{B}{2} - C\right)\right]$$
with $\tilde{B} = \sum_{s_i \in S} \mu_i^2 / \sigma_i^2$ and $\tilde{C} = \sum_{s_i \in S} c_i \mu_i^2 / \sigma_i^2$.

Examples of continuous variables derived directly from GPS trajectories would be travel time and effort to/from a cell, demographic attributes such as average age or income in a cell, "survival" or waiting time in a cell, etc.

In addition to the distributions listed above, we have also implemented a fully nonparametric scan statistic which learns the relevant quantiles of historical counts.

In either case, the likelihood ratio is then computed for a massive set of regions that (i) match the shape and size of the clusters we are interested in detecting and (ii) which densely cover the considered space. The scanning window is either an interval (in time), a circle or an ellipse or a rectangle (in space) or a cylinder with a circular or elliptic or rectangular base (in space-time). Multiple different window sizes are used. The window with the maximum likelihood is the most likely cluster, $S_{most \ likely \ cluster} = \operatorname{argmax}_{S} LR(S)$, that is, the cluster least likely to be due to chance. We perform statistical significance testing by randomization resulting in a p-value assignment to this cluster.



Fig.2: Illustration of the scanning procedure.

When there are multiple clusters in the data set, the secondary clusters are either evaluated as if there were no other clusters in the data set or –optionally- adjusted for other clusters in the data in the following iterative manner: At every iteration only the most likely cluster is reported, subsequently removed from the data set, including all cases and controls (Bernoulli model) in the cluster while the population (Poisson model) is set to zero for the locations and the time period defining the cluster. Then, a completely new analysis is conducted using the remaining data. This procedure is then repeated until there are no more clusters with a p-value less than a user specified maxima or until a user specified maximum number of iterations have been completed, whichever comes first.

Additionally, CitySense offers the following choices regarding secondary clusters are available to the user:

(i) No Geographical Overlap: Secondary clusters will only be reported if they do not overlap with a previously reported cluster.

(ii) No Cluster Centers in Other Clusters: While two clusters may overlap,

there will be no reported cluster with its centroid contained in another reported cluster.

(iii) No Cluster Centers in More Likely Clusters: Secondary clusters are not centered in a previously reported cluster.

(iv) No Cluster Centers in Less Likely Clusters: Secondary clusters do not contain the center of a previously reported cluster.

(v) No Pairs of Centers Both in Each Others Clusters: Secondary clusters are not centered in a previously reported cluster that contains the center of a previously reported cluster.

(vi) No Restrictions = Most Likely Cluster for Each Grid Point: The most extensive option is to all present clusters in the list, with no restrictions. This option reports the most likely cluster for each grid point. This means that the number of clusters reported is identical to the number of grid points.

In a final step, CitySense augments and filters the space-time clusters in a hybrid fashion that reflects the particular application to nightlife.

1) We are usually only interested in clusters with a *substantially* increased rate. Hence, regardless of its statistical significance, we only report clusters where the relative risk is above a context specific threshold, $q > q_c > 1$.

2) While the scan statistic elegantly overcomes the *multiple hypothesis testing* dilemma, some applications do not have to or aspire to correctly adjust for multiple testing. In those situations. one should not fix the overall false alarm rate but allow it to grow with the number of cells under surveillance. For example, a nightlife recommendation system should report local hot spots in a city regardless of their adjusted significance which implicitly depends on the size of the grid. We address this issue by reporting - in addition to the significant space-time clusters - any region with a score higher than some fixed threshold.

<u>Time Component</u>: The above outlined algorithm identifies only spatial clusters, the extension to space time clusters is straightforward. In our application, we are mainly interested in "live" clusters, i.e. clusters that have emerged within some time interval *T* and are still present. If we denote the present time by $t_{present} = 0$, then we wish to find spatial regions S with higher than expected counts/measures than expected during the entire time interval (-T,0). In our terminology, these would be named "prospective clusters". The expression for the likelihood ratio is very similar to the one given above, except that now the baseline, over dispersion and the counts are time dependent, denoted by $\mathbf{b_i}(t)$, $\mathbf{o_i}(t)$ and $\mathbf{c_i}(t)$ and the products are taken over all spatial locations $s_i \in S$ and all times $-T \le t \le 0$:

$$LR(S,T) = \frac{\max_{q>1} \prod_{s_i \in S, -T \le t \le 0} \Pr[c_i(t) \sim NB(q \cdot b_i(t), o_i(t))]}{\prod_{s_i \in S, -T \le t \le 0} \Pr[c_i(t) \sim NB(b_i(t), o_i(t))]}$$

We typically discretize time into 10 - 30 minute chunks and maximize the likelihood ratio over space and time. The nightlife recommendation system usually explores clusters that persist up to a few hours.

Adjustment for covariates:

We assume that for each cell we have a historical time series of counts or measurements as well as external covariates such as weather, holidays, extraordinary events, etc.

Depending on the length of this historical data set, one can estimate time varying periodic effects at various scales. In its most basic form time is discretized into bins such as day-of-week, hour-of-day and month-of-year and we fit a generalized linear model to the observed counts/ measurements in each cell using standard regression software such as R or Matlab. Since this model attempts to learn the "normal" behavior as a function of the above listed predictors, we typically remove any data from the training set that are known to belong to unusual events such as holidays and other anomalies. Figure 3 shows an example of an average week hour pattern for a particular grid cell.



Fig.3: Visualization of expected counts as a function of hour-of-day (horizontal axis) and day-of-week (vertical axis). To guide the eye, we added the marginal distributions.

More sophisticated modeling strategies involve generalized additive models that estimate both shape and location parameters. The advantage being that time is modeled in a continuous fashion and interactions such as seasonality – day of week can be adjusted for. Figure 4 shows an example of a week hour pattern along with its fluctuations for a particular grid cell. Note that the zero value of the week hour is anchored at midnight, Sunday.



Fig.4: Mean and empirical quantiles of observed counts inside a particular grid cell as a function of weekhour. Overlaid are the corresponding Poisson quantiles. Note the overdispersion on Sat and Sun midday, which may be due to the inherent larger variability of human activity on weekends

4. Outlook:

When you use Citysense, the application learns about the kinds of places you like to go from GPS – without ever sharing that information. In its next release, Citysense will not only tell you where everyone is right now, but where everyone like YOU is right now. The application will compare your history and preferences with those of other users, and show you where you're most likely to find people with similar tastes at that moment. So each person's nightlife map will look a little different, and will display a unique top hotspot list. That's why we save your location when you use Citysense: to remember what you like. Of course, you don't have to keep a personalized nightlife profile. You can delete your data from our system anytime you want. You created your data: you own it. But showing up in Chicago for the first time and seeing the top places you're likely to find people with similar tastes as yourself at midnight – that's pretty useful.

We are serious about privacy and data ownership. Citysense already contains billions of location data points, which it uses to identify nightlife hotspots. When someone opens Citysense, the program references their current location to better understand the city's nightlife – with total anonymity. We never share your location, ever. We don't collect email addresses or phone numbers. We don't use passwords. In fact, we have a revolutionary new data ownership policy wherein people actually own any information they create. Citysense is opt-in, all the time. Anything Citysense collects, users can delete. You'll find the delete button easily accessible whenever you open the program. To read more, see our Principles.

References:

- Kulldorff M, Nagarwalla N. Spatial disease clusters: Detection and Inference. Statistics in Medicine, 14:799-810, 1995.

- Kulldorff M. A spatial scan statistic. Communications in Statistics: Theory and Methods, 26:1481-1496, 1997

- Kulldorff M. Prospective time-periodic geographical disease surveillance using a scan statistic. Journal of the Royal Statistical Society, A164:61-72, 2001.

- Daniel Neill and Andrew Moore, Anomalous Spatial Cluster Detection, Proceedings of the KDD 2005 Workshop on Data Mining Methods for Anomaly Detection

- Daniel Neill, PhD thesis, Carnegie Mellon University (2006).

- Real-time Outbreak and Disease Surveillance (RODS) system <u>http://www.rods.pitt.edu/</u>

- D. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. Annals of Statistics, 32(3):962–994, 2004

- P.A. Rogerson, Monitoring Point Patterns for the Development of Space-Time Clusters, J. Royal Statistical Society A, Vol. 164, No. 1, pp. 87-96.

- J. M. Loh, Z. Zhu, Accounting for spatial correlation in the scan statistic, The Annals of Applied Statistics. Volume 1, (2007), 560-584.