Spatio-temporal variations in the urban rhythm: the travelling waves of crime
Supplementary Material

Marcos Oliveira[1*], Eraldo Ribeiro[2], Carmelo Bastos-Filho[3], Ronaldo Menezes[1]

1 GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany
2 Computer Science, Florida Institute of Technology, Melbourne, Florida, USA
3 Universidade de Pernambuco, Recife, Pernambuco, Brazil
* `moliveira@biocomplexlab.org`

Contents

1 Data

In our work, we used official data of cities with respect to (1) their criminal activities, (2) their geography, and (3) their resident population. Here we summarise the data sources and the preprocess steps performed in order to carry out our experiments.

1.1 Criminal events

We analysed data sets of criminal occurrences in disaggregated level that contains the longitude and latitude of each offence. We obtained this data from 12 cities from United States, retrieved from the respective police offices of each considered city via their websites that are described in Table 1.

Though these data sets have their own particularities, each criminal event in any of them is characterised by the following:

- `type` - the category of the criminal event;
- `address` - the address where the crime occurred;
- `location` - the latitude and longitude where the crime occurred.
- `date` - when the offence happened.

Note that the `date` field may present different granularity such as the day including hour or the part of the day (e.g., morning, evening). In our work, however, we aggregate the number of offences that happened in each week.

Moreover, police offices may employ different terms when referring to certain categories of crime. They may also include subcategories of a type of crime (e.g., theft of bicycle, pickpocket). In our work, we focus on thefts. To analyse 'theft' in general, we grouped together events in each city which are described by different terms but are still theft; for this, we used as a guide the definitions from FBI[1]. That is, we renamed the type of crime in each record to 'theft'. Table 2 presents the terms that are used by each police office and that we grouped to perform our analysis. Note that some cities have only information of the broad term (i.e., 'theft').

Table 1: Data source for each considered city.

City	Crime data source
Atlanta	http://www.atlantapd.org/crimedatadownloads.aspx
Chicago	https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2
Hartford	https://data.hartford.gov/Public-Safety/Police-Incidents-01012005-to-Current/889t-nwfu
Kansas City	https://data.kcmo.org/browse?q=crime&Type=[object%20Object]&sortBy=relevance&utf8=%E2%9C%93
New York	https://data.cityofnewyork.us/Public-Safety/Historical-New-York-City-Crime-Data/hqhv-9zeg
Philadelphia	https://www.opendataphilly.org/dataset/crime-incidents
Portland	http://www.civicapps.org/datasets
Raleigh	https://data.raleighnc.gov/Police/Police-Incident-Data-from-Jan-1-2005-Master-File/csw9-dd5k
San Francisco	https://data.sfgov.org/Public-Safety/Map-Crime-Incidents-from-1-Jan-2003/gxxq-x39z
Santa Monica	https://data.smgov.net/Public-Safety/Police-Incidents/kn6p-4y74
Seattle	https://data.seattle.gov/Public-Safety/Seattle-Police-Department-Police-Report-Incident/7ais-f98f
St. Louis	http://www.slmpd.org/Crimereports.shtml

Table 2: The terms grouped for theft in each considered data set.

City	Terms selected for thefts
Atlanta	'LARCENY-FROM VEHICLE', 'LARCENY-NON VEHICLE'
Chicago	'THEFT'
Hartford	'06* -LARCENY'
Kansas City	'stealing from bldg', 'Stealing from Auto', 'Stealing From Auto', 'Stealing All Other', 'Stealing from Buildi', 'Stealing Auto Parts / ', 'stealing', 'Stealing from Bldg', 'Stealing Auto Parts', 'Stealing Pickpocket', 'stealing ACC', 'stealing all other', 'Stealing From Buildi', 'stealing-bldg', 'Stealing ACC', 'stealing acc', 'Stealing other', 'stealing oth', 'STEALING', 'Stealing Purse Snatc', 'stealing from auto', 'stealing accessories', 'Stealing from buildi', 'stealing from build', 'STEALING ACC'
New York	'GRAND LARCENY'
Philadelphia	'Thefts', 'Theft from Vehicle'
Portland	'Larceny'

[1] https://ucr.fbi.gov/crime-in-the-u.s/2010/crime-in-the-u.s.-2010/violent-crime/violent-crime

Raleigh	'LARCENY (CIVILIAN USE ONLY)', 'LARCENY / FROM BUILDING (-$50)', 'LARCENY / POCKET-PICKETING / FELONY (-$50)', 'LARCENY / PURSE-SNATCHING / FELONY (-$50)', 'LARCENY / FROM MOTOR VEHICLE / FELONY (-$50)', 'LARCENY / PURSE-SNATCHING / FELONY ($50-$199)', 'LARCENY / FROM BUILDING ($50-$199)', 'LARCENY / PURSE-SNATCHING (-$50)', 'LARCENY / POCKET-PICKING ($200-$1,000)', 'Larceny / Pocket-Picking', 'LARCENY / ALL OTHERS ($200-$1000)', 'LARCENY / ALL OTHERS ($1000+)', 'LARCENY / FROM BUILDING / FELONY ($50-$199)', 'Larceny / Theft from Building', 'Larceny / All Other', 'LARCENY / MOTOR VEHICLE PARTS / ACC / FELONY($200-1000)', 'Larceny / Purse-Snatching', 'LARCENY / FROM MOTOR VEHICLES / FELONY (OVER $1,000)', 'LARCENY / ALL OTHERS ($50-$199)', 'LARCENY / PURSE-SNATCHING / FELONY ($200-1,000)', 'LARCENY / MOTOR VEHICLE PARTS / ACC ($200-$1,000)', 'Larceny / Theft from Motor Vehicle', 'LARCENY / FROM MOTOR VEHICLES ($50-$199)', 'LARCENY / PURSE-SNATCHING ($50-$199)', 'LARCENY / FROM BUILDING / FELONY (OVER $1,000)', 'LARCENY (NO LONGER USED)', 'LARCENY / PURSE-SNATCHING ($200-$1,000)', 'LARCENY / FROM BUILDING / FELONY (-$50)', 'LARCENY FROM MOTOR VEH (NO LONGER USED)', 'LARCENY / MOTOR VEHICLE PARTS / ACC / FELONY (OVER $1,000)', 'LARCENY / FROM MOTOR VEHICLES / FELONY ($200-$1,000)', 'LARCENY / MOTOR VEHICLE PARTS / ACC / FELONY (-50)', 'LARCENY / MOTOR VEHICLE PARTS / ACC ($50-$199)', 'LARCENY / POCKET-PICKING / FELONY ($200-$1,000)', 'LARCENY / ALL OTHERS / FELONY(-$50)', 'LARCENY / ALL OTHERS (-$50)', 'LARCENY / PURSE-SNATCHING / FELONY (OVER $1,000)', 'LARCENY / POCKET-PICKING (-$50)', 'LARCENY / POCKET-PICKING ($50-$199)', 'LARCENY / FROM MOTOR VEHICLES ($200-$1,000)', 'LARCENY / FROM MOTOR VEHICLES (-$50)', 'LARCENY / FROM BUILDING / FELONY ($200-$1,000)', 'LARCENY / MOTOR VEHICLE PARTS / ACC / FELONY ($50-$199)', 'LARCENY / MOTOR VEHICLE PARTS / ACC (-$50)', 'LARCENY / FROM MOTOR VEHICLES / FELONY ($50-199)', 'LARCENY / POCKET-PICKING / FELONY (OVER $1,000)', 'LARCENY / POCKET-PICKING / FELONY ($50-$199)', 'LARCENY / FROM BUILDING ($200-$1,000)', 'LARCENY / ALL OTHERS / FELONY($50-$199)', 'Larceny / Theft of MV Parts-Accessories', 'LARCENY / ALL OTHERS / FELONY($200-$1000)'
San Francisco	'LARCENY / THEFT'
Santa Monica	'Larceny -Purse-snatch', 'Larceny -Pickpocket', 'Larceny -General', 'Larceny -Vehicle Parts / Acc', 'Larceny -From Building', 'Larceny -Other', 'Larceny -From Vehicle'
Seattle	'THEFT-AUTO PARTS', 'THEFT-PRSNATCH', 'THEFT-OTH', 'THEFT-CARPROWL', 'THEFT-BUILDING', 'THEFT-PKPOCKET', 'THEFT-BICYCLE', 'THEFT-AUTOACC'
St. Louis	'LARCENY-PICKPOCKET $500-$24,999', 'LARCENY-MTR VEH PARTS OVER $25,000', 'LARCENY-FROM MTR VEH OVER $25,000', 'LARCENY-ALL OTHER UNDER $500 / ATTEMPT', 'LARCENY-PURSESNATCH UNDER $500 / ATTEMPT', 'LARCENY-FROM BUILDING $500 -$24,999 / ATTEMPT', 'LARCENY-ALL OTH / FRM PRSN / $150-$199.99', 'LARCENY-ALL OTHER $500 -$24,999', 'LARCENY-FROM MTR VEH $500 -$24,999', 'LARCENY-FROM MTR VEH UNDER $500 / ATTEMPT', 'LARCENY-MTR VEH PARTS $500 -$24,999', 'LARCENY-PICKPOCKET UNDER $500 / ATTEMPT', 'LARCENY-FROM BUILDING UNDER $500 / ATTEMPT', 'LARCENY-MTR VEH PARTS UNDER $500 / ATTEMPT', 'LARCENY-PURSESNATCH UNDER $500', 'LARCENY-ALL OTH / FRM PRSN / UNDER $500', 'LARCENY-FROM BLDG $200-$749.99', 'LARCENY-ALL OTHER OVER $25,000', 'LARCENY-PICKPOCKET UNDER $500', 'LARCENY-ALL OTHER / $150-$199.99', 'LARCENY-FROM MTR VEH UNDER $500', 'LARCENY-PURSESNATCH $500-$24,999', 'LARCENY-FROM BUILDING $500 -$24,999', 'LARCENY-ALL OTHER UNDER $500', 'LARCENY-FROM BUILDING UNDER $500', 'LARCENY-FROM BUILDING OVER $25,000', 'LARCENY-MTR VEH PARTS UNDER $500', 'LARCENY-FROM BLDG $150-$199.99', 'LARCENY-PICKPOCKET OVER $25,000'

1.2 Geospatial information

To analyse criminal events across the city, we used official geospatial data to create the shapefiles of the cities. For each city, we used the boundaries of the respective U.S. state from the U.S. Census Bureau[2] and the TIGER (Topologically Integrated Geographic Encoding and Referencing) shapefiles with granularity of blocks (delimited in 2010). To have the geography of the considered cities in the study, we clipped each shapefile with the bounding box of each city, using the bounding boxes retrieved from the OpenStreetMap initiative[3].

[2] https://www.census.gov/
[3] http://nominatim.openstreetmap.org/

Note that to carry out spatial analyses, we have to project each crime data set on the same spatial projection of their respective boundaries. This procedure was not needed in most of the cities, since the crime data sets and boundaries shared the same spatial reference `EPSG:4326`. Still, we needed to perform this procedure for the data of two cities, as described in Table 3.

Table 3: The spatial references in the crime data sets and the shapes of the locations.

Location	Crime data sets	Boundaries
St. Louis	EPSG:2815	EPSG:4326
Portland	EPSG:2269	EPSG:4326
Other cities	EPSG:4326	EPSG:4326

1.3 Population

To analyse spatial heterogeneity in cities, we split each city into regions that have approximately same number of people. For this, we need data regarding the spatial distribution of the population in each city. We gathered data about the total resident population in the smallest spatial units available of the considered locations from official census. In our work, we used the 2010 census data from the U.S. Census Bureau[4] that provides the total population (variable name `P1`) in block level.

2 Splitting cities

In this work, we analysed the time series of crime occurring in small spatial units across a city. We constructed these spatial units using a method developed by Oliveira et al. [1] which divides a city into regions of approximately same resident population. In summary, this approach first creates a graph based on the spatial and census data in which each node of the graph represents the same amount of population. With this representation of the city, we can use a partitioning algorithm to divide the graph into parts of same number of nodes. These parts also have approximately same total population, and thus we can use them as our regions. For more details, see [1].

2.1 The number of splits and the analysis of crime

To analyse the waves of crime across a city c, we now have to choose the number of regions R_c that the city will be divided using the aforementioned approach. The value of R_c has to be chosen in such way that the aggregation level leads to units that represent the city. For this, we first investigate the crime rate of the regions while splitting each city. We count the number of regions $R_c^\varphi(r)$ that exhibit crime rate higher than φ when the city c is composed by r regions of same population size. Fig. 1 shows $R_c^\varphi(r)$ with $\varphi = 1.0$ as we increase r for all considered cities. We found that $R_c^\varphi(r)$ increases with r until it reaches a maximum value $R_c^\varphi(r_c^u)$ of regions. Note that this result is expected, given that crime is unevenly distributed: as we divide a city into an increasing number of regions, the spatial units decrease their area which implies lower probability that offences occur in the same unit.

In this work, we want to investigate the waves of crime across the city and the ideal is to have in hands (1) broad coverage of the city and (2) time series that have data. Therefore, in our analysis of small spatial units, we use the maximum number of regions to have enough data points in the time series. That is, we examine each city using r_c^u regions while setting $\varphi = 1.0$, depicted in Fig. 1 with the vertical lines.

Note that each city c can be divided into R_c regions in different ways or arrangements that depends on the partitioning algorithm used to split the graph (i.e., the representation of the city), as briefly described in the beginning of Section 2. We followed Oliveira et al.[1] and used a stochastic partitioning

[4] https://factfinder.census.gov/

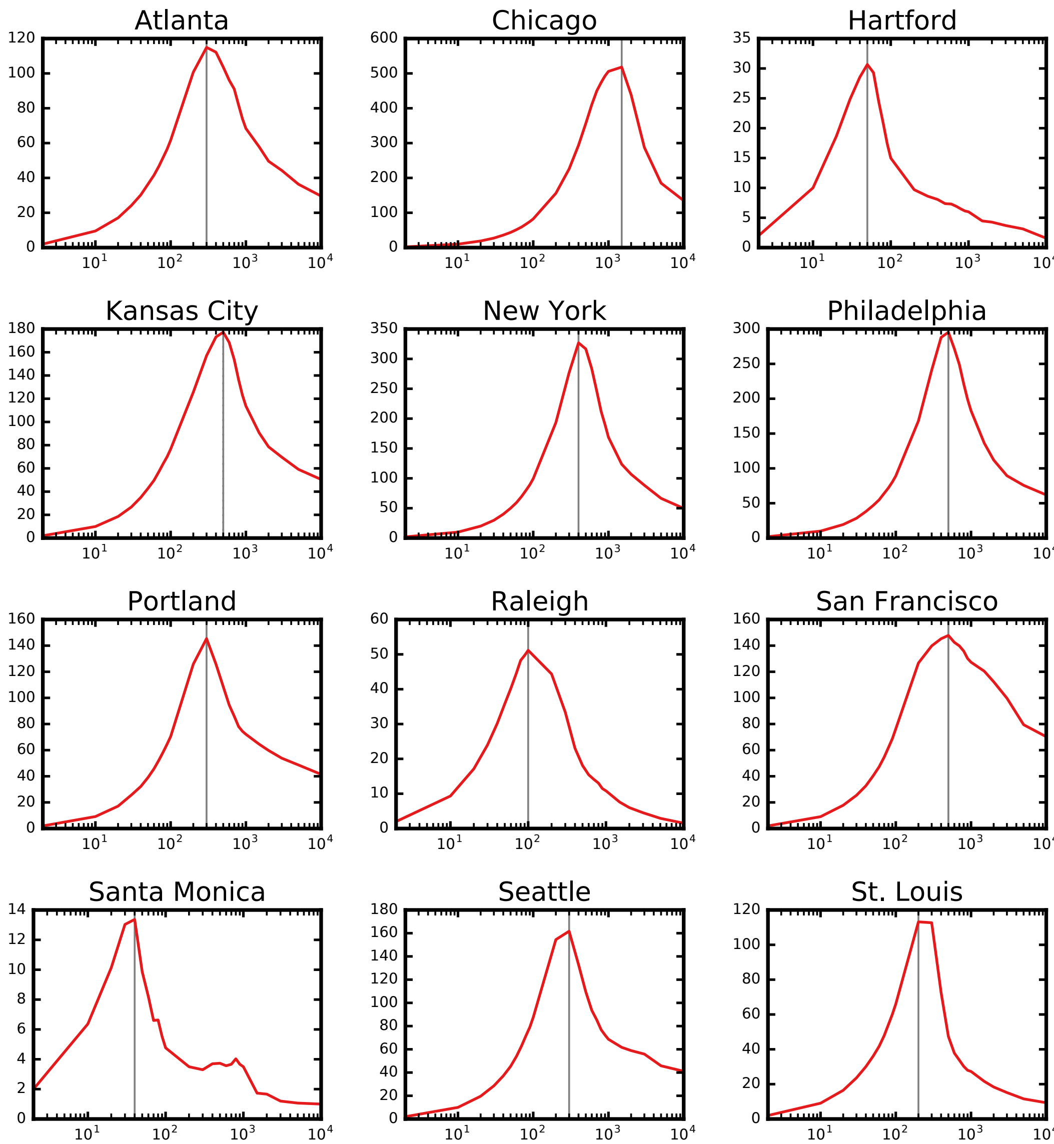

Figure 1: The number of regions with at least one crime per week, namely R^{φ}, increases as we divided the city until R^{φ} reaches the maximum value when the city is composed by r_u^c regions.

algorithm for splitting the cities. With an stochastic algorithm, we can generate different divisions of the city. For this, we use the KaFFPa (Karlsruhe Fast Flow Partitioner) algorithm to partition each city (i.e., its graph representation) using different seeds for the random number generator [2]. For each city c we generated 30 different arrangements in which each one comprises of R_c same-population divisions of the city. With each of these arrangements, we can build the time series of crime occurring in each region of the city.

3 Preprocessing time series

To analyse the criminal series, we preprocess the raw data to (1) decrease skewness in the data, (2) remove trends, and (3) decrease intra-month variance. We created the time series $r(t)$ using seven-days time windows, which means that each data point $r(t')$ is the number of occurrences in that particular week t'. For the sake of simplicity, this notation regards to both city-level and local-level time series.

The skewness in criminal data may occur due to different reasons such as crime sprees and crime repeats which leads to temporal crime concentration [3]. In criminal studies, researchers tend to decrease this skewness by applying transformations on the data [4, 5, 6]. In our analysis, we use a log-transformation to decrease the skewness as the following:

$$x(t) = \log_{10}[r(t) + 1]. \tag{1}$$

To remove the long-term trends, we first use the moving average of a series, defined as:

$$\mathrm{M}^{n_1,n_2}[x(t)] = \frac{1}{n_2 - n_1} \sum_{n=j_1}^{n_2} x(t+j), \tag{2}$$

to determine the long-term tendency in the series, using $n_1 = -26$ and $n_2 = 26$ (i.e., one year). Then we remove this trend from the series as the following:

$$d(t) = x(t) - \mathrm{M}^{-26,26}[x(t)], \tag{3}$$

thus $d(t)$ consists of the detrended time series of crime in a city.

In this work, we are interested on the cycles that are higher than one-month period. However, the high variance between weeks in each month might hide such tendencies in the series, thus we apply the moving average filter with window size equals to 5 in order to remove any intra-week dynamics:

$$y(t) = \mathrm{M}^{0,5}[d(t)]. \tag{4}$$

These preprocessing steps are shown for the considered time series in Fig. 2.

4 Wavelet analysis and composed analysis

In this work, we examine the wavelet transform of the time series, defined as the following for a given discrete sequence $Y = \{y(1), y(2), \ldots, y(N)\}$:

$$\mathrm{W}_Y(s,n) = \sqrt{\frac{\delta t}{s}} \sum_{t=1}^{N} y(t)\, \psi\left[\frac{(t-n)\,\delta t}{s}\right], \tag{5}$$

where δt is the uniform step between the observations of Y and $n \in \{1, 2, ..., N\}$.

Specifically, we use the *local wavelet spectrum* as a tool to evaluate the periodicity in crime, defined as:

$$|W(s,n)|^2. \tag{6}$$

If we average the local wavelet spectrum across time, we have the so-called *global wavelet spectrum*, defined as

$$\overline{W}^2(s) = \frac{1}{N} \sum_{n=1}^{N} |W(s,n)|^2, \tag{7}$$

which gives us the periods present in the time series, similarly to Fourier spectrum. The *scale-averaged wavelet power*, defined as

$$\overline{W}^2_{j_1,j_2}(n) = \frac{\delta j \delta t}{C_\delta} \sum_{j=j_1}^{j_2} \frac{|W(s_j,n)|^2}{s_j}, \tag{8}$$

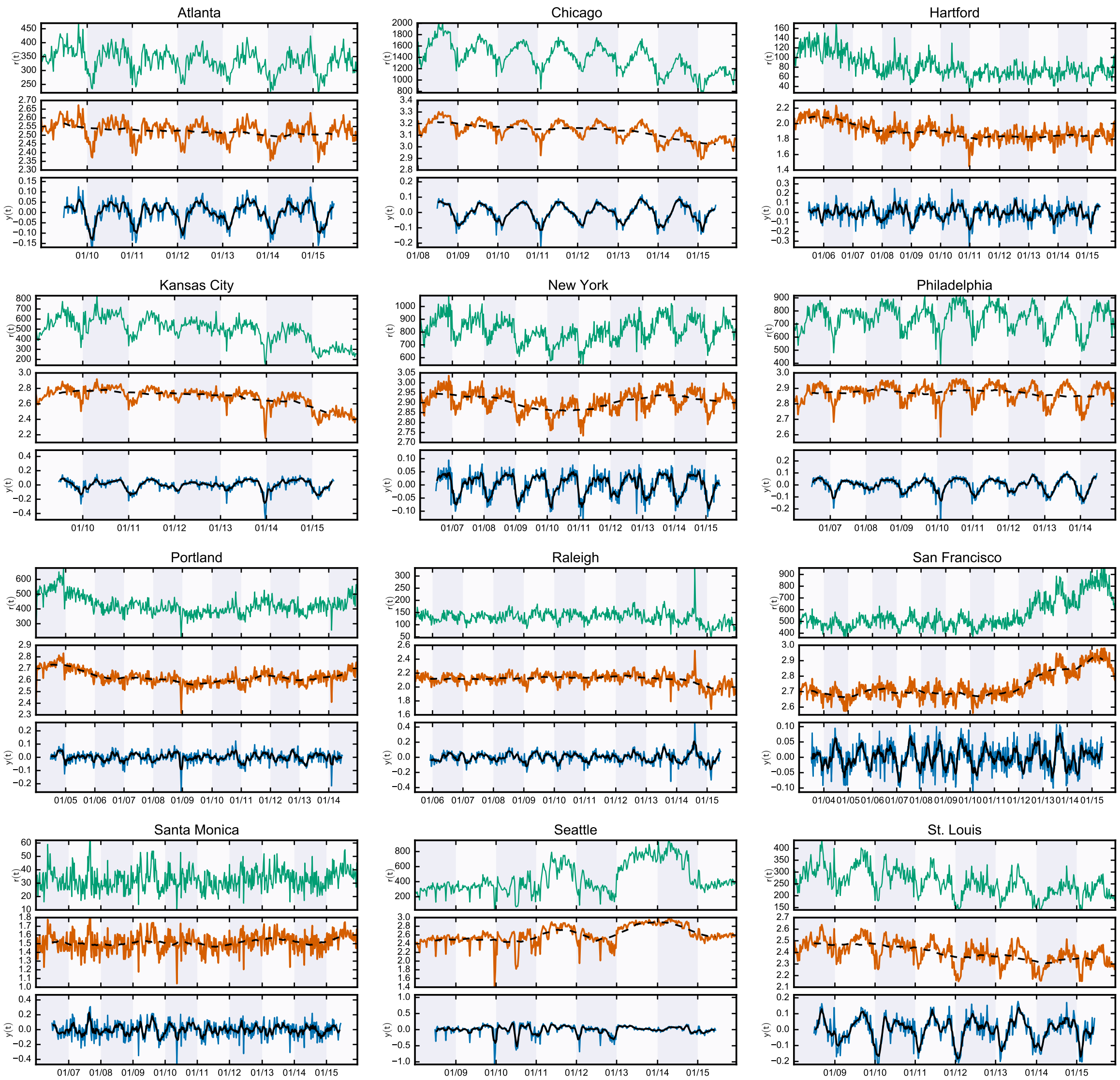

Figure 2: Preprocessing the time series of theft. First, we apply a log transformation (Eq. 1), then we remove the trend in the theft time series (Eq. 3), then finally smooth the data using moving average (Eq. 4).

enables us to analyse the temporal evolution of a periodic signal in terms of a given band (j_1, j_2). In Eq. 8, C_δ is the reconstruction factor defined for each wavelet function calculated by reconstructing a delta function δ from its wavelet transform. We used the values found in [7] for C_δ with respect to the Morlet wavelet with $\omega_0 = 6$, that is, $C_\delta = 0.776$. Following also [7], we use

$$s_j = s_0 2^{j\delta j} \quad \text{and} \quad J = \frac{1}{\delta j} \log_2 \left(\frac{N \delta j}{s_0} \right),$$

for $j = \{0, 1, \ldots, J\}$, where s_0 is the smallest resolvable scale [7].

With the wavelet power spectrum, we have a measure of local variance, so we are also interested on its statistical significance. For this, we used the method developed by Torrence and Compo [7]. We tested the wavelet power against a null model that generates a background power spectrum P_k, given by:

$$D\left(\frac{|\overline{W}_X(s,n)|^2}{\sigma_X^2} < p\right) = \frac{1}{2} P_k \chi_\nu^2, \tag{9}$$

where $\nu = 2$ for complex wavelets (our case) and $\nu = 1$ for real-valued wavelets [7]. In this work, we use red noise with 1-lag autocorrelation $\alpha = 0.72$ as the null model. In Fig. 3, we show the global wavelet spectrum of the wavelet transform of the city-level time series of each city against the null model. Fig. 4 depicts this test for the scale-averaged power for $b = (0.8, 1.1)$ for all considered cities, where the horizontal lines designates the 95% confidence for red noise for each time series.

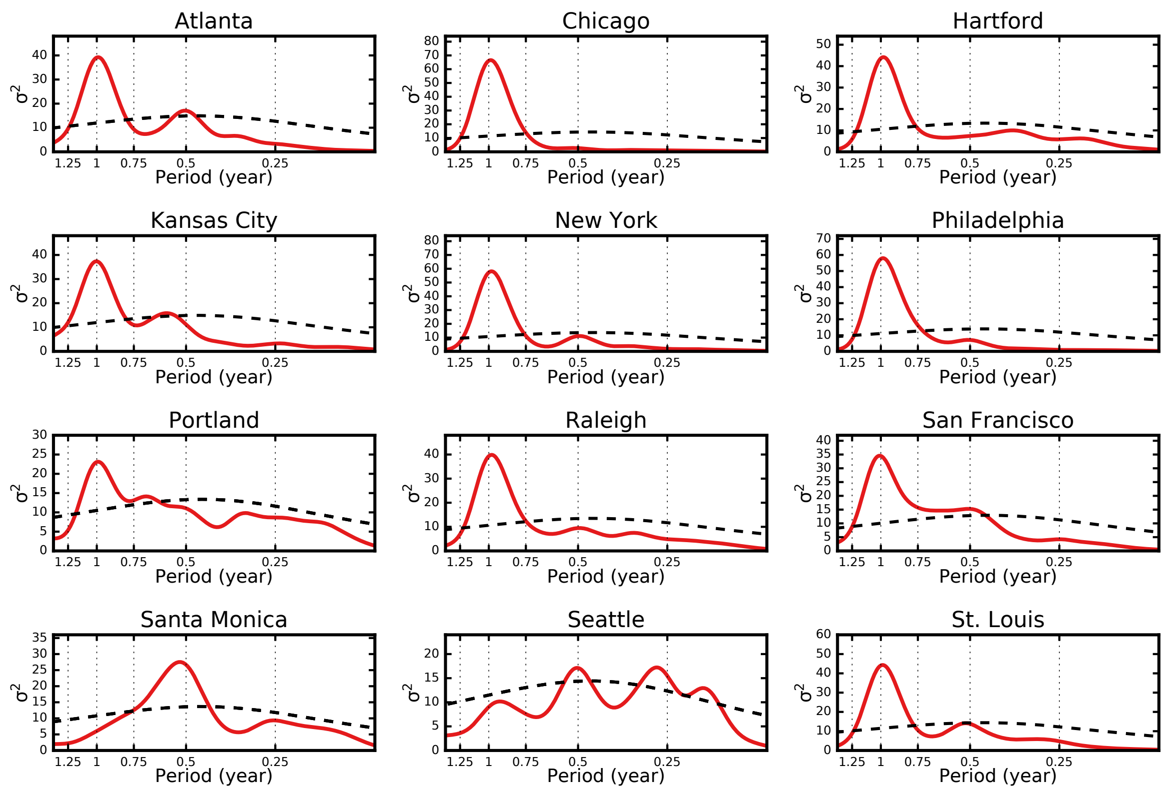

Figure 3: The global spectrum of the wavelet transform. Most of the cities exhibit circannual period of crime, except for Santa Monica and Seattle which exhibit a semestral period. The dashed line is the null model based on red noise.

To take into account the heterogeneity across the city, we examined the time series in small spatial units (i.e., regions) in the cities. We proposed the *composed spectra* $C_c(s)$, defined as the number of regions exhibiting time series with statistically significant global spectrum at each period s in a city c. To build $C_c(s)$, we counted the number of regions $N_c(s)$ with significant period s, then divided $N_c(s)$ by the total number of regions in the city. $C_c(s)$ is shown in Fig. 5 for all cities. We also proposed the *composed scale-averaged power* $C_c^b(t)$ that is defined as the number of regions that exhibit a statistically significant band $b = (j_1, j_2)$ at the time step t in the city c. Specifically, given a city c, we counted the number of regions that have a statistically significant band at each time step t. Fig. 8 depicts $C_c^b(t)$ for all cities.

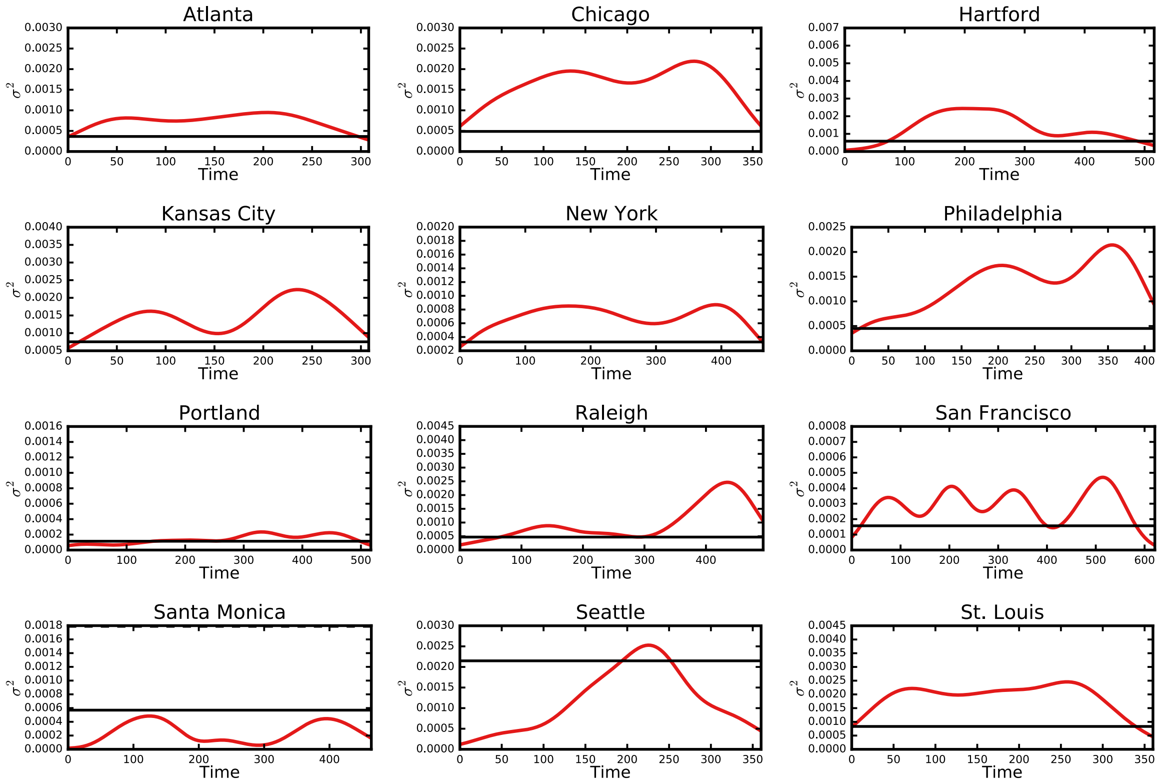

Figure 4: The scaled-averaged power of the wavelet transform. Most of the cities present statistically significant 1-year component throughout the time series. That is, most of the cities exhibit stationary circannual rhythm of crime.

5 Fitting Δt_b^c

To describe the mobility of the criminal waves, we examined the random variable Δt_b^c, defined as the amount of time that a Y_i^c (i.e., a region in the city c) exhibits a significant periodicity with respect to the band b. For each city, we measured Δt_b^c using the circannual band, as shown in Fig. 9. In practical terms, we counted the number of time steps that the scale-averaged power is continuously significant in each region for a given city. We found that the probability distribution of Δt_b^c (w.r.t. circannual waves) decays much earlier than the total time of the criminal series. To select the model for Δt_b^c distribution, we followed the procedures described by Clauset et al. [8], and compared the following distributions: truncated power law (TP), lognormal (LN), exponential (EX), and stretched exponential (SE). For this task, we used the Python library `powerlaw` [8, 9]. As also shown in Fig. 9, the stretched exponential distribution gives a good fit when compared to the other distributions.

References

[1] Marcos Oliveira, Carmelo Bastos-Filho, and Ronaldo Menezes. The scaling of crime concentration in cities. *PLOS ONE*, 12(8):e0183110, aug 2017.

[2] Peter Sanders and Christian Schulz. Think Locally, Act Globally: Highly Balanced Graph Partitioning. In *Proceedings of the 12th International Symposium on Experimental Algorithms (SEA'13)*, volume 7933 of *LNCS*, pages 164–175. Springer, 2013.

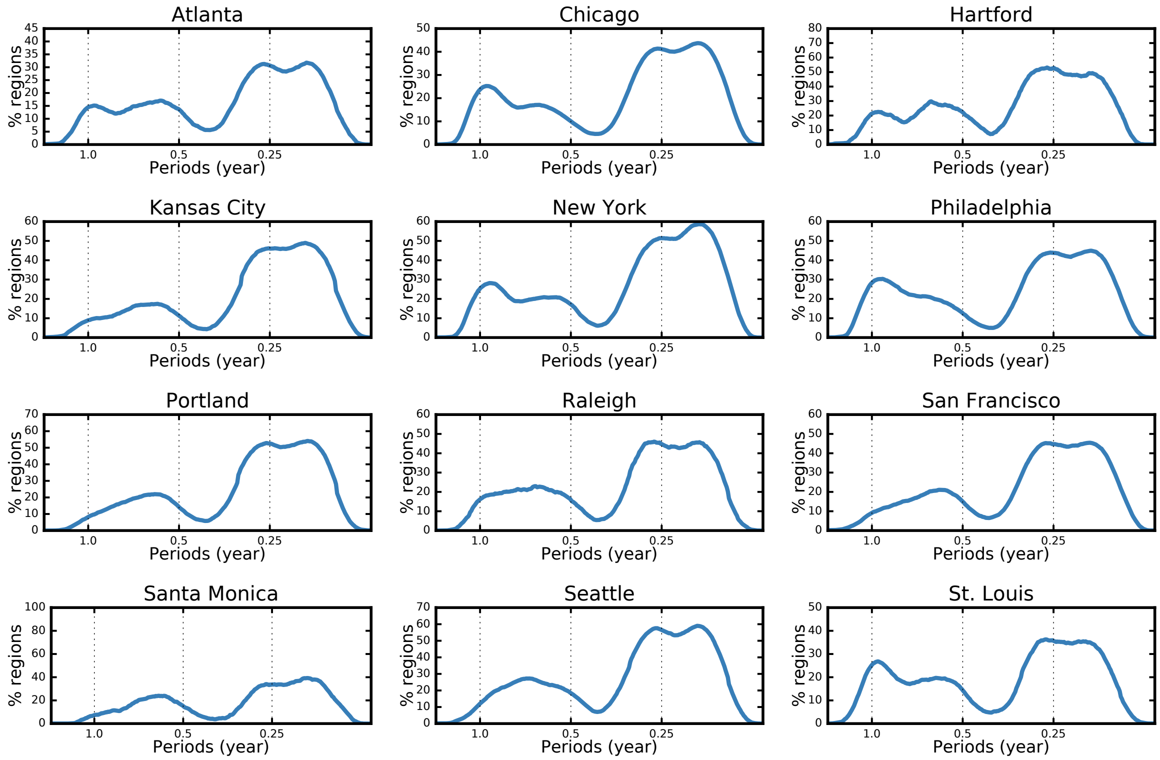

Figure 5: The composed spectra presents a holistic perspective of the periodicity in the city while accounting for the heterogeneity. Crime exhibits a signature in the periods in that the three months is the leading period.

[3] Graham Farrell. Crime concentration theory. *Crime Prevention & Community Safety*, 17(4):233–248, nov 2015.

[4] John R. Hipp, Patrick J. Curran, Kenneth A. Bollen, and Daniel John Bauer. Crimes of Opportunity or Crimes of Emotion? Testing Two Explanations of Seasonal Change in Crime. *Social Forces*, 82(4):1333–1372, jun 2004.

[5] Ellen G. Cohn and James Rotton. Even criminals take a holiday: Instrumental and expressive crimes on major and minor holidays. *Journal of Criminal Justice*, 31(4):351–360, jul 2003.

[6] Ellen G. Cohn and James Rotton. Weather, Seasonal Trends and Property Crimes in Minneapolis, 1987–1988. A Moderator-Variable Time-Series Analysis of Routine Activities. *Journal of Environmental Psychology*, 20(3):257–272, sep 2000.

[7] Christopher Torrence and Gilbert P Compo. A Practical Guide to Wavelet Analysis. *Bulletin of the American Meteorological Society*, 79(1):61–78, jan 1998.

[8] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. Power-law Distributions in Empirical Data. *SIAM Review*, 51(4):661–703, nov 2009.

[9] Jeff Alstott, Ed Bullmore, and Dietmar Plenz. powerlaw: a python package for analysis of heavy-tailed distributions. *PloS one*, 9(1):e85777, 2014.

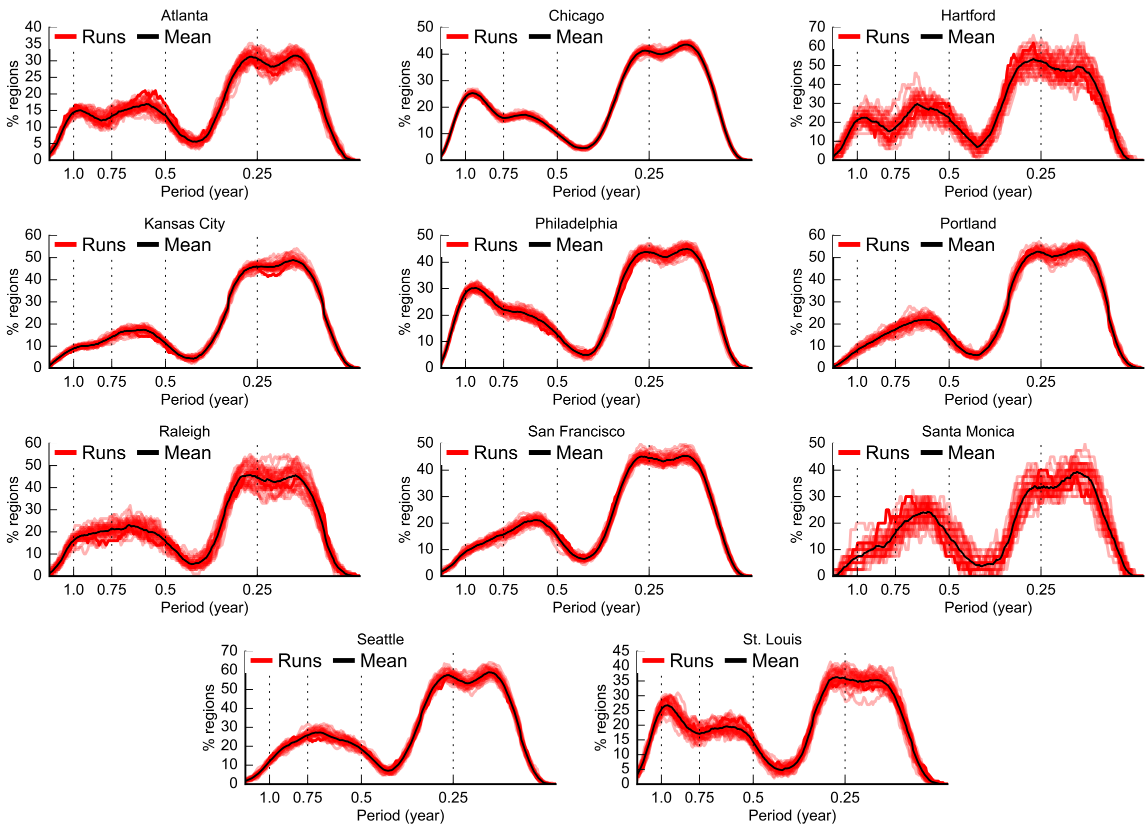

Figure 6: Splitting each city multiple times. With 30 different arrangements, we compute their composed spectra. The characteristic curve does not change much with different ways to split the city.

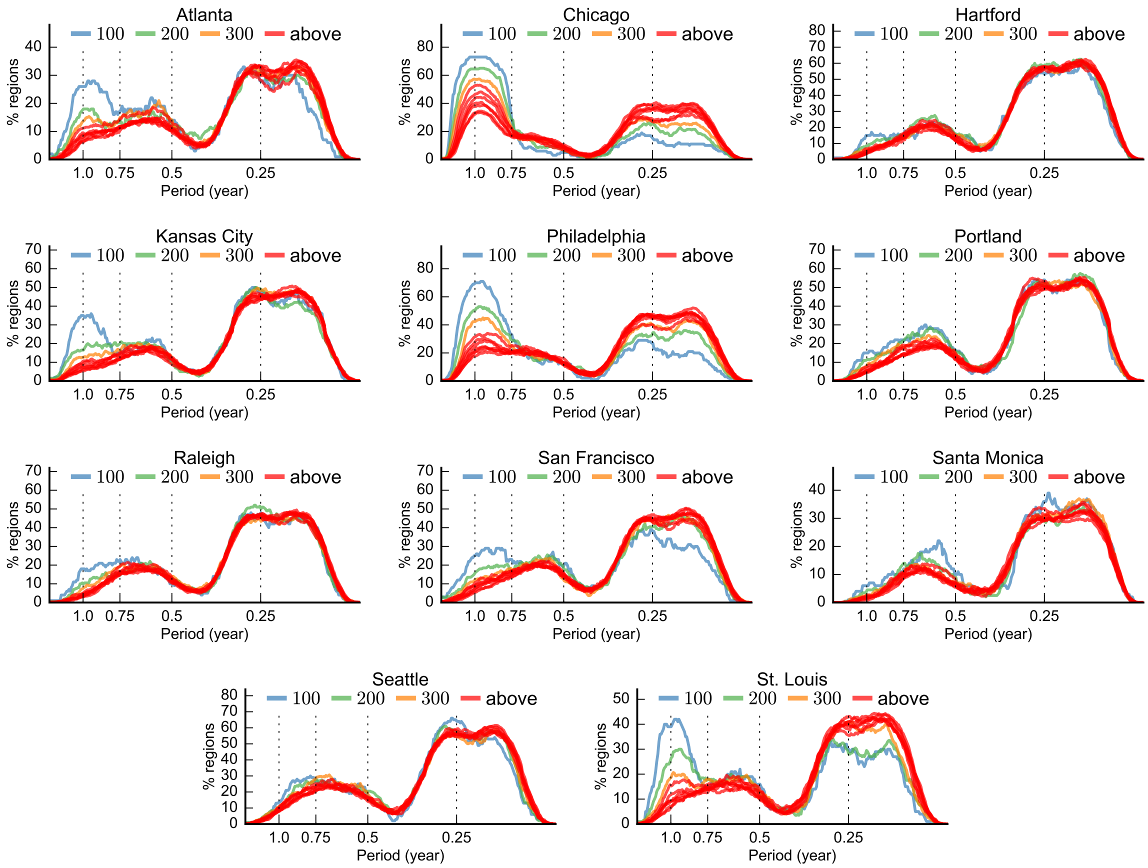

Figure 7: Splitting each city into an increasing number of regions. The composed spectra after splitting each city with an increasing number of regions (100, 200, 300, 400, 500, 600, 700, 800, 900, and 1000). The composed spectra converge to a characteristic curve after a certain number of regions.

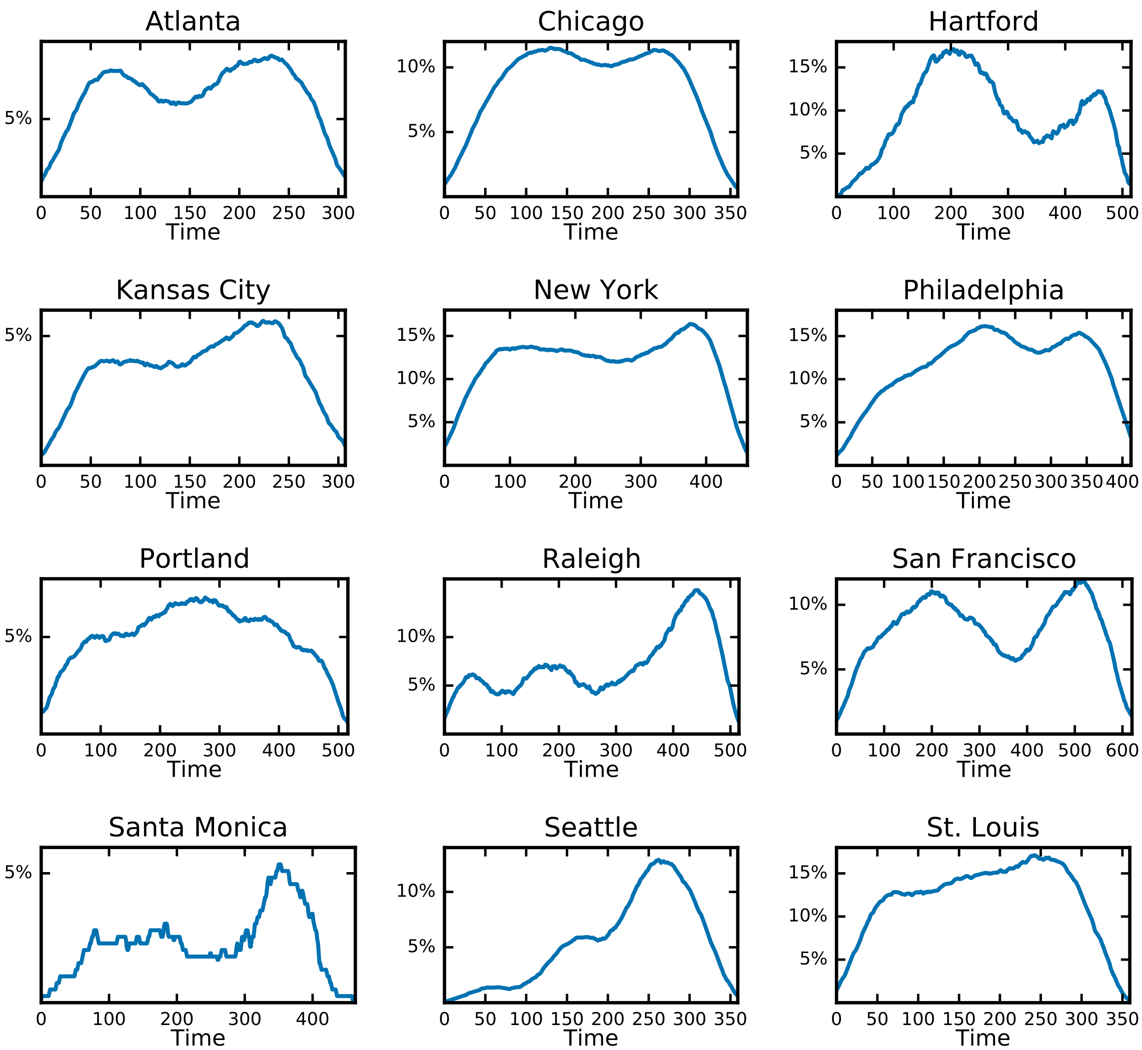

Figure 8: The composed scale-averaged power of the circannual band shows the dynamics in the whole city with respect to the 1-year wave. In most of the cities, the amount of regions with this periodicity keeps fairly the same.

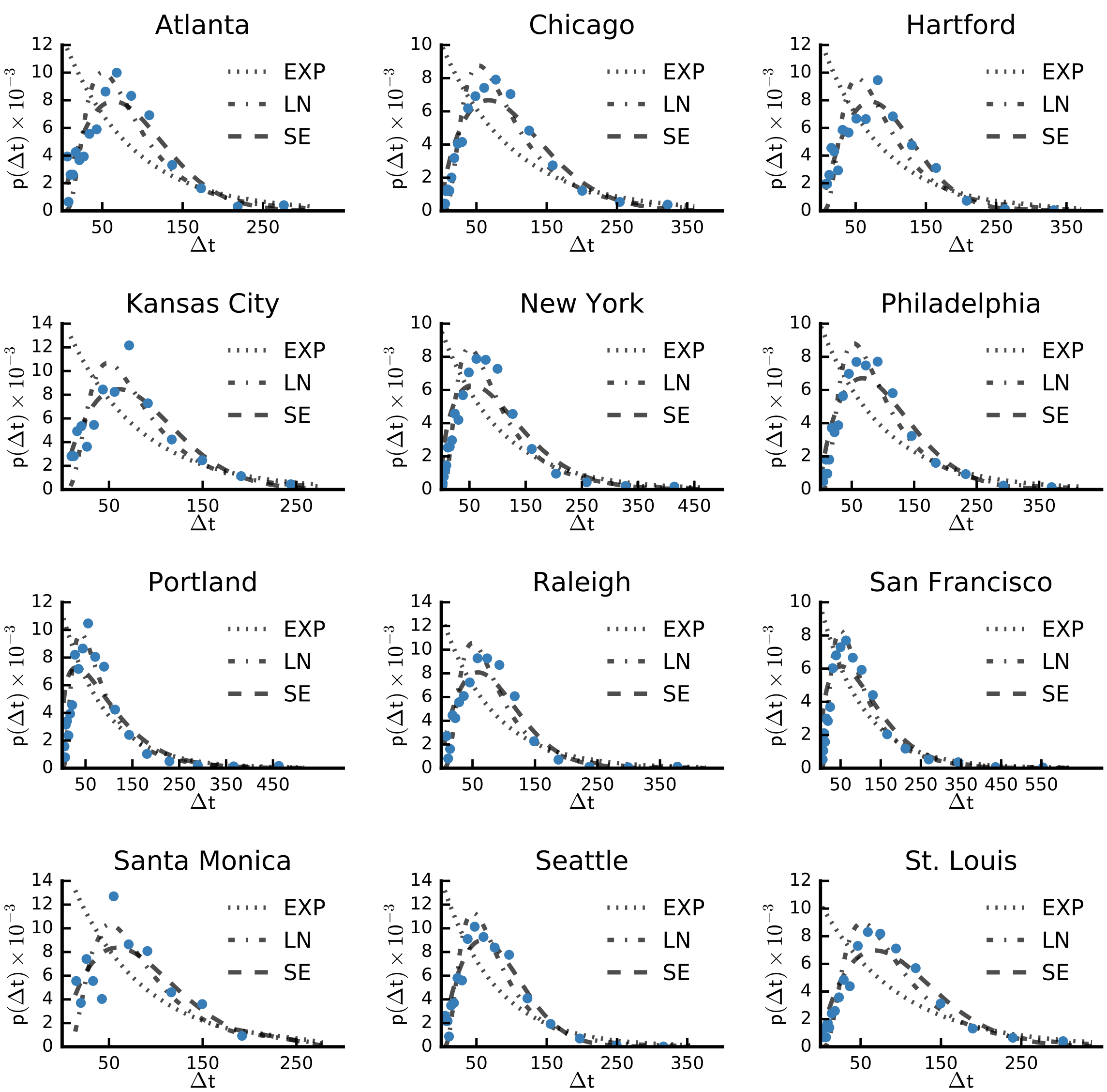

Figure 9: The amount of time Δt a region exhibits the circannual period tends to be smaller than the total amount of data. The stretched exponential distribution (SE) gives a good fit for $p(\Delta_t)$ when compared to exponential (EXP) and log-normal (LN) distributions.