Better burst detection in two dimensions

Christopher Cameron

April 20, 2016

Abstract

Two dimensional shifted wavelet trees (SWT) produce better bounding boxes for concentrations of points within point clouds. This report describes the issues inherent with fixed grid burst detection systems and evaluates the performance of an SWT based burst detection approach against a fixed grid approach. Finally, I evaluate the consequence of various inclusion criteria on the number of users unambiguously assignable to a single home box. An implementation of a SWT based algorithm for proposing home-box-candidates is available on the project git repository at `github.com/georgeberry/educ/tree/master/twitter/better_boxes`

1 Introduction

Given a collection of longitude and latitude pairs for a user, we seek to identify the bursts or concentrations of points in two dimensions. Concentrations tend to be contained within an approximately 100 by 100 meter square. Recognizing that the commonly used fixed-grid approach: [1] is theoretically inappropriate, [2] is technically incorrect with the potential to miss the largest point clouds, [3] creates an issue with multiple candidates that requires post-processing in about half the cases and [4] wastes API calls, I propose an alternative accumulation method that is designed for burst detection and provably capable of detecting arbitrarily placed bursts.

2 Shortcomings of fixed grid burst detection

The fixed grid approach described in the literature uses a single grid with dividing lines placed at intervals of 0.001 degrees of longitude and latitude (approximating a 100 by 100 meter square area). The number of points within each box is used to rank the boxes and the box(es) with the highest counts are taken to contain the highest concentration of points. When used for burst detection, fixed grid approaches yield the counts within the units of a specific division of space. If the divisions represent cities, counties or states, the counts within the division have a natural interpretation. When the divisions are arbitrary and small, as is the case with 100 by 100 meter boxes, the count within the boxes holds no particular meaning.

Fixed grid approaches are used to sample the distribution of objects in space and can used to estimate the density and size of the object population. Fixed

grid counting techniques have also been employed to detect areas of high concentration of objects (refs). The issue of detecting spatial concentration is fundamentally different from sampling grid points for population inference. The point of burst detection is not to count number of objects withing a specific division of space. Rather, the purpose is to find the division of space that contains the concentration of points.

Despite its popularity in the literature, fixed grid counting is inappropriate for spatial burst detection. The arbitrary placement of the dividing lines can create counting issues for point clouds that are not centered at coordinates evenly divisible by .0005 degrees latitude and longitude. These off-center point clouds spill into neighboring grid cells, diluting the count (Figure 2). In the worst case, a point cloud divided over 4 cells would appear to be as significant as a well centered point cloud that was 4 times smaller.

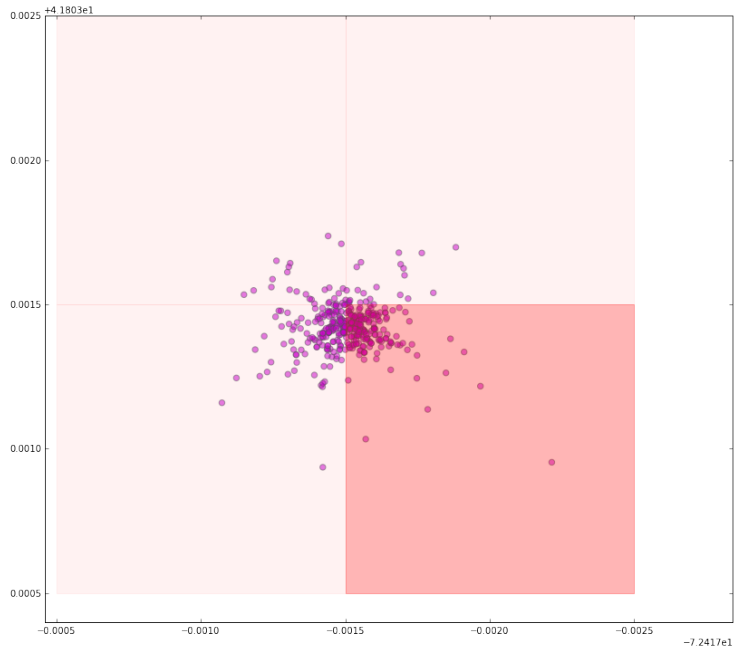

Figure 1: A failure of fixed-grid counting. If a concentration of points spans the boundaries of a fixed grid, the point cloud is distributed into the neighboring boxes (light pink). The darker box contains the most points, but the points in the box are not a good summary of the total point cloud and the size of the maximal box does not reflect the size of the point cloud. This concentration could be overlooked in favor of a better-centered but smaller point cloud.

3 Burst detection with shifted wavelet trees

Zhu and Shasha [2003] introduced a *Shifted Wavelet Tree* data structure that supports sliding window aggregation at different scales and guarantees detection of any burst whose size exceeds a specified threshold within a specified interval width (major grid size). In practice, the data structure is easily implemented

by using a smaller sub-grid of size of interval/n and then aggregating over all possible contiguous n by n sub-grid regions to identify candidates.

Aggregation results in a collection of candidate major-grid boxes that can be ranked by size. The highest ranked candidate is selected as the location of the regional burst and all overlapping candidate boxes are discarded. This still allows for adjacent boxes if the points are truly distributed in a bi-modal pattern with center-of-mass separation greater than the major grid size.

The proofs in Zhu and Shasha [2003] show that only one sub-division is necessary to guarantee the detection of bursts at the scale of the major grid. Using smaller sub-grid regions yields box cells with somewhat better flexibility to reach additional points. The tradeoff for more flexible boundaries is a greater number of candidate cells to sort and moving from 2 to 10 sub-boxes considerably increases the run time.

Figure 2 shows the result of using a shifted wavelet tree (SWT) over the same set of points shown in Figure 1. With the flexibility of SWT, the highest ranked bounding box contains most of the points. The number of home-box-candidates was reduced from four to one box, potentially reducing the number of API calls by a factor of 4.

Using more than two subdivisions (Figure 3 is not necessary to reliably detect bursts but does allow more flexibility to pick up additional points.

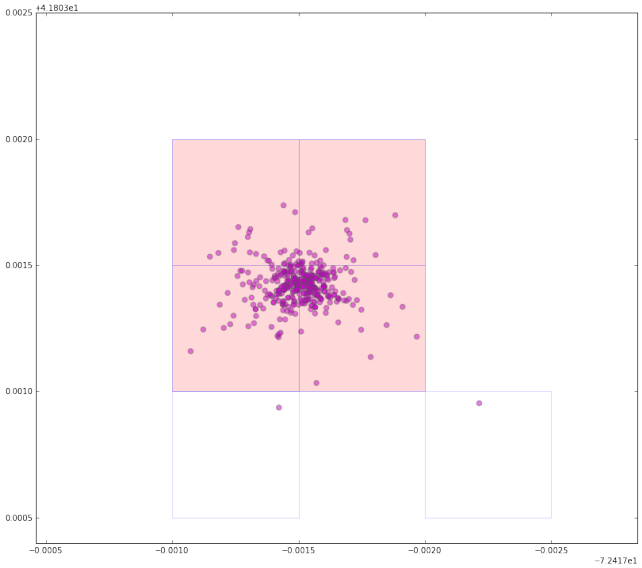

Figure 2: Shifted wavelet trees with half-width grids support offsets of .0005 degrees in each dimension. The additional flexibility allows the algorithm to consider and select the best of several candidate squares. The blue boxes represent occupied sub boxes and the larger pink square represents the best large-grid square produced by aggregating over all 2x2 sub-box sets.

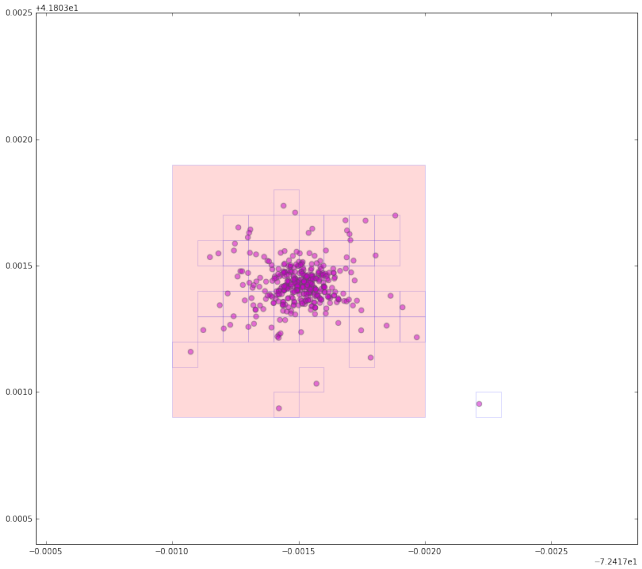

Figure 3: Using 10 divisions within the grid lets the final candidate box shift south slightly to encompass an extra point. The blue boxes represent occupied sub boxes and the larger pink square represents the best large-grid square produced by aggregating over all 10x10 sub-box sets.

4 Performance

The performance of SWT based approaches can be evaluated in terms of computational demand and the quality of the output.

4.1 Computation

Shifted wavelet trees cost additional computing time because they consider many more candidate grid cells. Table 1 shows that the 2-sub-box approach adds a modest amount of computing time. The benefit of using more than two subdivisions is unlikely to be worth the computation cost.

Divisions	Run Time (minutes)	Performance
0-subdivisions (simple)	129	1
2-subdivisions	173	1.34
10-subdivisions	1892	14.6

Table 1: Computing candidate boxes for 50k users with geocoded tweets. Using two sub-boxes provides much better bounding boxes with an acceptable 30% increase in runtime.

4.2 Improvement in home box assignments

Comparing the candidate boxes proposed by the simple and the 2-sub box SWT approach demonstrated in Figure 2, I find the following benefits:

A 5% increase in the number of users with at least one home box. These users had point clouds distributed over a few boxes and none of these boxes satisfied the 10-unique-day-cutoff. The shifted box was able to accommodate these point clouds and satisfy the 10-day criteria.

A 24% increase in the number of users with exactly one box. A little less than half are people moving from 0 to 1 box and the rest are people with point clouds like those in Figure 1, who moved from 2-4 boxes under the simple box approach to a single box in the shifted wavelet tree approach.

5 User capture rates

This section provides some details about the result of attempting home box assignment to approximately fifty thousand users. About 70% of users do not have enough geo-coded tweets to generate home-box candidates.

Among 50,540 U.S. users with any geotagged tweets, 18,841 users have one or more candidate home boxes that meet the 10-day criteria. About half the users have exactly one candidate box and the other half have two or more candidate boxes. The boxes can be ranked by the number of weekend night tweets in each box. When the count of the weekend evening tweets in the highest ranked box is greater than double the size of the next largest box, we assume the largest box is the home location.

The 2x criteria applied to the 18,841 users with any home boxes produces 15788 users with a single unambiguous home box.

It is possible for the number of evening tweets in a qualified box to be quite small, even if the number of unique days on which the user tweeted from within the box exceeds the criteria. It would make sense to restrict our consideration to boxes with more than a few evening tweets since the signal of 3-4 tweets is pretty weak. Table 2 shows the consequence of imposing minimum weekend evening cell size criteria on the number of users with any boxes and the number of users that can be unambiguously assigned to a box using a 2x criteria.

Minimum number of weekend evening Tweets	Users with any candidate boxes	Users with unambiguous assignment to single box
0	18841	15788
5	15557	13457
10	12466	10909
20	8976	8004
50	4797	4456

Table 2: Assignment of home boxes for 50k users with geocoded tweets. Overall, between 20 and 30% of users can be unambiguously assigned to a single home box. Most users do not have enough data for assignment.

In addition to filtering out boxes with weak home-box signals, we might require that a home box contain a sizable fraction of all evening tweet activity. If a home box candidate has, for instance, 20 tweets but the total nighttime tweet activity is a widely dispersed cloud of thousands of points, then the home box candidate is not clearly a place of evening refuge. On the other hand, if

the user tweets less frequently in the evening, then the 20 point candidate box might represent ten or twenty percent of all the night time activity. A criteria based on the percent of weekday night time activity can be paired with the minimum weekday night time criteria to filter out the cases with the highest uncertainty. Table 3 shows that the percent-of-activity filter produces only marginal reduction in the number of users that can be assigned to home box.

		Percent of weekend evening tweets			
		0%	5%	10%	20%
Minimum number of weekend evening Tweets	0	15788	15738	15557	15006
	5	13457	13451	13401	13170
	10	10909	10908	10898	10763
	15	9211	9211	9205	9121
	20	8014	8014	8008	7938

Table 3: The number of users unambiguously assignable to a home box with two filtering systems in place. Enforcing a minimum percentage criteria, such that the home box candidate must have more than some minimum percentage of all weekday evening tweets in addition to some minimum number of weekend evening tweets filters out cases with more uncertainty.

6 Recommendations and observations

1. Use the shifted wavelet tree structure to avoid the counting issues inherent in a fixed grid approach to burst detection.

2. In addition to the ten-unique-day rule, disregard boxes with fewer than some minimum number of weekend evening tweets. A **minimum requirement of 10 weekend evening tweets** looks like an acceptable starting point, as it still results in about 20% of users assignable to home boxes while eliminating assignments based on weak signals.

3. When the largest candidate home box is more than twice as large as the next largest candidate, we may be able to accept the largest candidate as unambiguous. This should be evaluated against Minsu's test set to judge how often this criteria might miss a viable second candidate box. The relative size of best to second best ranked candidates is a reasonable way to choose between candidate boxes for users with multiple boxes. **Requiring that the first box be at least twice as large and the next best candidate is a conservative filter that allow unambiguous home box assignment for about half of the users with multiple boxes.**

4. Users with no clear home box among multiple candidate boxes account for about 10% of the users with boxes. If these users could be successfully related to a home value, we could expand the number of income labeled users by about 10%.

5. In addition to containing some minimum number of weekday evening tweets, acceptable candidate boxes should also contain some appreciable *proportion* of the user's weekday evening tweets. This filter does not

greatly reduce the number of cases, even with a relatively high criteria of 20%. Including this criteria improves the believably of all home box assignments without eliminating too many additional cases, so the **successful home box candidates should represent at least 20% of the weekend evening tweets.**

References

Yunyue Zhu and Dennis Shasha. Efficient elastic burst detection in data streams. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '03*, page 336. ACM Press, 2003. doi: 10.1145/956750.956789. URL http://www.scopus.com/inward/record.url?eid=2-s2.0-77952383186&partnerID=tZOtx3y1 http://portal.acm.org/citation.cfm?doid=956750.956789.