SANDIA REPORT

SAND2014-1825
Unlimited Release
Printed March, 2014

Tensor Analysis Methods for Activity Characterization in Spatiotemporal Data

Michael J. Haass, Mark H. Van Benthem, Edward M. Ochoa

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Approved for public release; further dissemination unlimited.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd.
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2014-1825
Unlimited Release
Printed March, 2014

Tensor Analysis Methods for Activity Characterization in Spatiotemporal Data

Michael J. Haass
Cognitive Systems

Mark Van Benthem
Computational Materials and Data Science

Edward Ochoa
Mission Engineering and Analysis

Sandia National Laboratories
P.O. Box 5800
Albuquerque, New Mexico 87185-1327

Abstract

Tensor (multiway array) factorization and decomposition offers unique advantages for activity characterization in spatio-temporal datasets because these methods are compatible with sparse matrices and maintain multiway structure that is otherwise lost in collapsing for regular matrix factorization. This report describes our research as part of the PANTHER LDRD Grand Challenge to develop a foundational basis of mathematical techniques and visualizations that enable unsophisticated users (e.g. users who are not steeped in the mathematical details of matrix algebra and mulitway computations) to discover hidden patterns in large spatiotemporal data sets.

ACKNOWLEDGMENTS

We wish to acknowledge Kristina Czuchlewski, Principal Investigator (PI) for the PANTHER LDRD Grand Challenge for her leadership, consultation and encouragement and Daniel Dunlavy (SNL Data Analysis and Informatics Department) for his consultation and support on Sandia's Tensor Toolbox.

CONTENTS

1. Introduction 7
   1.1. Background 7
2. Analysis of INFLUENZA OUTBREAKS 9
   2.1. Methodology 9
   2.2. Experimental 10
   2.3. Results 11
   2.4. Discussion 14
3. Analysis of Taxi Cab Activity Traces 15
   3.1. Methodology 16
   3.2. Experimental 17
   3.3. Results and Discussion 18
4. Conclusions 25
5. References 26
Distribution 28

FIGURES

Figure 1 Graphic representation of the bilinear data model. 8
Figure 2 Graphic representation of the trilinear, 3-way data model. 9
Figure 3. HHS Surveillance Regions 10
Figure 4. 3-factor NNTF model of influenza type A (H1) for years 1998-2012. 12
Figure 5. 4-factor NNTF model of influenza type A (H1) for years 1998-2012 13
Figure 6. 3-factor NNTF model of influenza type A (H3) for years 1998-2012. 14
Figure 7. 4-factor NNTF model of influenza type A (H3) for years 1998-2012. 14
Figure 8. Example GPS traces (in red) from one taxi in the San Francisco Bay Area. 16
Figure 9. Visualization of geospatial grid input data. 17
Figure 10. Factors 1 and 2 along the taxi fare dimension. 18
Figure 11. Geospatial patterns of activity described by factor 1 (without fare). 19
Figure 12. Geospatial patterns of activity described by factor 2 (with fare). 20
Figure 13. Zoomed view of factor 1 (without fare). 21
Figure 14. Factor 2 (with fare) along the geospatial dimension. 22
Figure 15. Factor 2 (with fare) along the geospatial dimension highlights a relatively high level of taxi service with fares on board from SFO to OAK. 23
Figure 16. Factors 1 (without fare) and 2 (with fare) along the temporal dimension. 24
Figure 17. Factors 1 (without fare) and 2 (with fare) along the temporal dimension when trip termination times are binned by day of week. 24

NOMENCLATURE

CDC	Center for Disease Control and Prevention
CRAWDAD	Community Resource for Archiving Wireless Data at Dartmouth
DOE	Department of Energy
DoD	Department of Defense
HHS	US Department of Health and Human Services
LDRD	Laboratory Directed Research and Development
NREVSS	National Respiratory and Enteric Virus Surveillance System
NNTF	Nonnegative tensor factorization
OAK	Oakland International Airport
PANTHER	Pattern Analytics To support High performance Exploitation and Reasoning
PCA	Principal component analysis
SFO	San Francisco International Airport
SNL	Sandia National Laboratories
WHO	World Health Organization

1. INTRODUCTION

Rapidly growing data sets from high-output remote sensing and reporting technologies are overwhelming DoD analysts and their production capabilities. New generations of sensors continually achieve increases in resolution (both spatial and temporal) and dimensionality (i.e. number of wavelengths). Although the current approach of hiring and training more analysts is typically used as a stop-gap measure, this human resource intensive approach to mitigate an ongoing, and growing, problem does not scale to the magnitude of data. Alternatively, computer-aided analytics may be developed to assist analysts with their mission data mining and reporting processes resulting in increased efficiency.

Tensor (also referred to as multiway array) factorization and decomposition offers unique advantages for activity characterization in spatio-temporal datasets because these methods are compatible with sparse matrices and maintain multiway structure that is otherwise lost in collapsing for regular matrix factorization. This report describes our research as part of the Pattern Analytics to Support High Performance Exploitation and Reasoning (PANTHER) Laboratory Directed Research and Development (LDRD) Grand Challenge to develop a foundational basis of mathematical techniques and visualizations that enable unsophisticated users (e.g. users who are not steeped in the mathematical details of matrix algebra and multiway computations) to uncover patterns in large spatiotemporal data sets by highlighting the appropriate dimensionality, or groupings, of the data.

As datasets grow in size, a human analyst's cognitive capacities such as working memory are exceeded, thereby increasing the difficulty of distinguish patterns from noise. Consequences of this increased difficulty may include a) failure to identify temporally stable patterns of activity or b) failure to identify rare, yet significantly anomalous activities. Multiway array methods can reduce the analyst's cognitive load and cue their attention by highlighting the appropriate reduced dimensionality or feature set to examine for maximal, or most relevant, information retrieval.

Two case studies were conducted to illustrate the applicability of the mathematical techniques to a broad range of spatiotemporal activity patterns and data types. The first case study examined and developed models of the extent and intensity of annual influenza outbreaks in the United States from region to region. The second case study examined and developed models of taxi cab activity in the San Francisco Bay Area.

1.1. Background

Multivariate statistical analysis methods are so-called "soft modeling" techniques that seek to decompose data into its simplest form. Principal component analysis (PCA) is one such modeling method that decomposes a matrix of data in a bilinear way subject to the constraints of columnwise orthogonality and ordered variance. For example, consider a collection of influenza data that represents the number of reported cases by month of the year for different geographical regions in the United States as a matrix **D**. Using PCA, one could decompose the data matrix as a set of scores, **T**, and loadings, **P**, that represent the temporal distributions and correlated regional distributions for the flu season.

$$\mathbf{D} = \mathbf{T}\mathbf{P}^{\mathrm{T}} \tag{0}$$

Since the scores in T are orthogonal and the loadings in P are orthonormal, the results can be difficult to interpret. Imposing alternative constraints on the bilinear model, such as nonnegativity, can generate results whose meaning may be easier to construe.

$$\mathbf{D} = \mathbf{X}\mathbf{Y}^{\mathrm{T}}, s.t. \mathbf{X} > 0, \mathbf{Y} > 0 \tag{0}$$

Here **X** is the matrix of temporal factors and **Y** is the matrix of region factors, all of whose elements are positive. Somewhat challenging in these models is the ambiguity in the results due to the rotation problem. In the absence of nonnegativity constraints, there are an infinite number of solutions to the right hand side of (0) all of which satisfy the least squares cost function equally. While imposing constraints on the **X** and **Y** factor matrices can limit the range of solutions, there is no guarantee that the solution one obtains is the most accurate representation of the underlying information. Figure 1 shows the bilinear data model in graphical form.

$$\mathbf{D} \approx \sum_{i=1}^{R} \mathbf{x}_i \mathbf{y}_i^{\mathrm{T}}$$

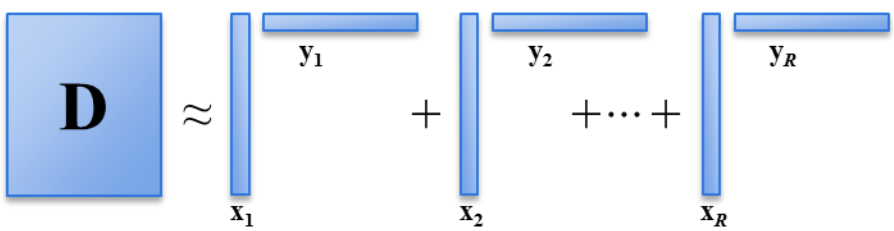

Figure 1 Graphic representation of the bilinear data model.

Fortunately, there are data types and analysis methods that can overcome these problems, even in the absence of constraints. Trilinear data and trilinear decomposition methods can relieve some of these problems. Trilinear data follow the model [1-6],

$$\mathfrak{D} = \otimes(\mathbf{X},\mathbf{Y},\mathbf{Z}) \tag{0}$$

where $\mathfrak{D}$ is now a 3-way array, or tensor, and the operator $\otimes$ represents the triple product of the pure component factor matrices **X**, **Y**, and **Z**. The solutions to the factor matrices can be estimated in a number of ways involving iterative methods and direct methods[7, 8]. A common, reliable, iterative method known as parallel factors analysis [1, 9, 10] or PARAFAC uses an alternating least squares scheme similar to MCR (multivariate curve resolution). While MCR alternately estimates **X** given **Y**, and then **Y** given **X** until convergence; PARAFAC-ALS (alternating least squares) estimates **X** given **Y** and **Z**, then **Y** given **X** and **Z**, and finally **Z** given **X** and **Y**. This approach lends itself well to our established expertise and fast algorithms developed for use with MCR. By imposing nonnegativity constraints on PARAFAC-ALS, the resulting factors are all nonnegative producing a nonnegative tensor factorization. Figure 2 portrays the trilinear model in graphical form.

$$\mathcal{D} \approx \sum_{i=1}^{R} \mathbf{x}_i \circ \mathbf{y}_i \circ \mathbf{z}_i$$

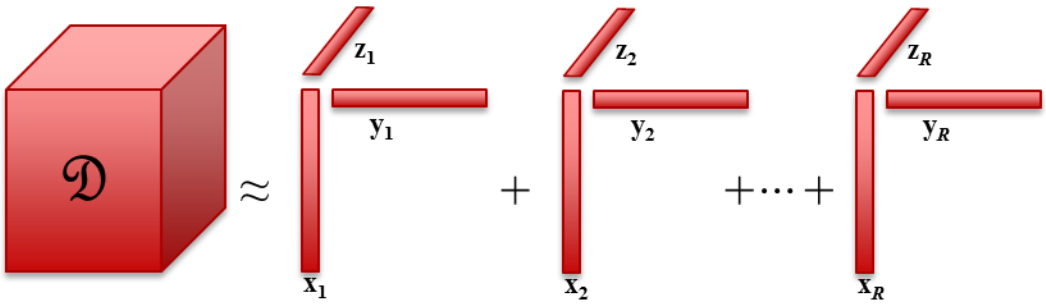

Figure 2 Graphic representation of the trilinear, 3-way data model.

For additional information on tensor decompositions and their applications, Kolda and Bader provide an excellent overview in their 2009 review article [11].

2. ANALYSIS OF INFLUENZA OUTBREAKS

Influenza outbreaks occur on a regular basis in the United States and around the world. The extent and intensity of these epidemics vary from year to year and region to region. Understanding the spread of influenza and modeling outbreaks can help medical professionals select treatment options and mitigation strategies [12]. The most common model in use is the Susceptible-Infectious-Recovered (SIR) model [12-15], which is a parametric model [16] involving terms of group size; probability of infection β; and most critical to epidemic prediction, the Basic Reproduction Number (R0). R0 describes the average number of new infections that each existing infection generates. Its determination leads to the estimation of the probability of infection. While the SIR model appears often in the literature as the epidemic model of choice, it can be too rigid to accurately model historical influenza data [15]. Statistical models that make use of actual epidemic data can provide historic insight into trends of the spread of influenza and, potentially, yield predictive behavior on burgeoning outbreaks.

2.1. Methodology

In this work, we consider influenza data collected over a number of years, from 1998 to 2012. Data were downloaded from the Center for Disease Control and Prevention (CDC) and grouped by season, geographical region and flu variant type. The data were sourced from the CDC website gis.cdc.gov/grasp/fluview/fluportaldashboard.html, which were derived from flu monitoring at approximately 80 collaborating U.S. World Health Organization (WHO) laboratories and 60 National Respiratory and Enteric Virus Surveillance System (NREVSS) laboratories. These facilities report the number of respiratory specimens tested and the number that test positive for influenza types A and B each week are reported to the CDC. In addition, most of the U.S. WHO collaborating laboratories report the influenza-A subtype (H1 or H3; where alpha, H, is the haemagglutinin antigen and the number, 1-12, is the subtype) of the viruses they test and the age or age group of the persons from whom the specimens were collected. The majority of NREVSS laboratories do not report subtype. Reports from these sources are combined and the weekly total number of positive influenza tests are collected in the weekly influenza update known as FluView [17]. Data are grouped by year, week, and region. The US Department of

Health and Human Services (HHS) specifies ten U.S. regions for influenza tracking. Figure 3 is a color-coded map that shows the ten regions in the U.S. and its territories and possessions.

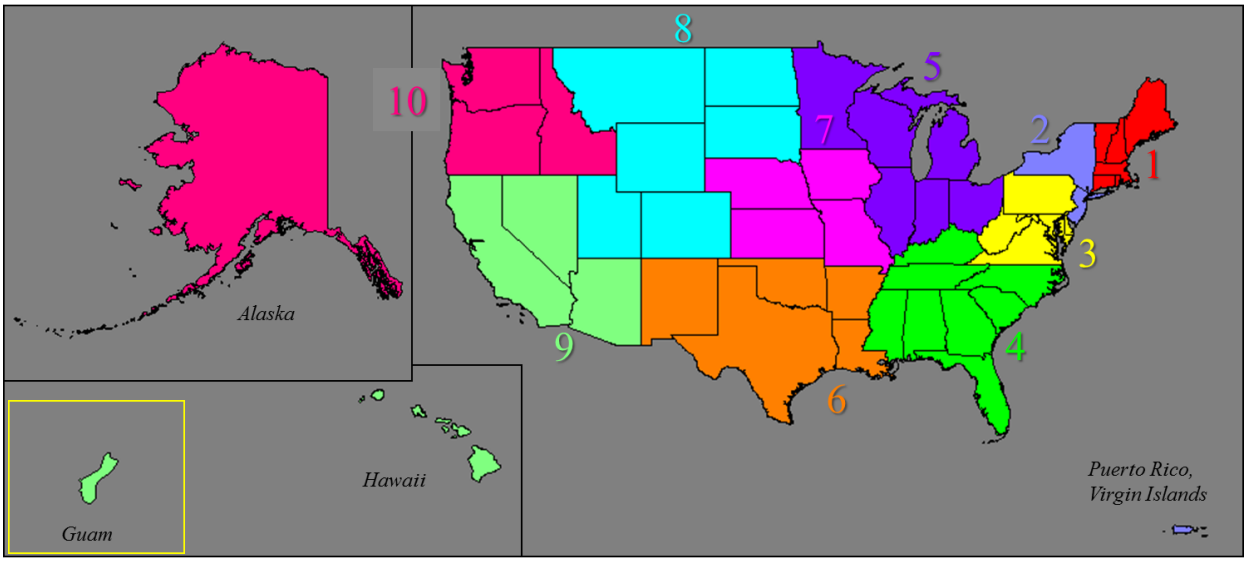

Figure 3. HHS Surveillance Regions

For this analysis, the data were grouped by year from 1998 to 2012. Years run from week 40 in the calendar year to week 39 of the following year (i.e., week 40 of 1997 to week 39 of 1998). The data were further grouped by week in the year. In the data, there are typically 52 weeks per year, but some have 53 due to the fractional number of weeks per calendar year. Existent week 53 data were added to week 52 data to accommodate these non-standard events. In the years 1998-2002 data for weeks 21-39 is nonexistent. In these cases, missing data was assigned a value of zero. Although this is not ideal, it allows analysis for all years as a first approximation. Finally, using principles from information theory [18, 19], the data were scaled using an entropy transformation in the form below in Equation 4.

$$d_{ijk} = \left(d_{ijk} + 1\right)\log\left(d_{ijk} + 1\right) \tag{0}$$

2.2. Experimental

The procured influenza data were preprocessed as described above organizing them as a week by year by region array with respective dimensions of 52x15x10. The data were then scaled using the entropy transformation. The preprocessed data were analyzed using nonnegative tensor factorization (NNTF). NNTF was performed using a fast PARAFAC algorithm developed at Sandia National Laboratories (SNL) [20] while imposing nonnegative least squares for all variables in each mode for all iterations. The nonnegative least squares algorithm was also developed at SNL [21]. Each factorization was initiated using a set of variables selected from a uniform random distribution with values between 0-1. Initial factors were reinitialized if they were found to lead to poorly conditioned solution sets. Iterations were terminated using an automatic criterion that evaluated the difference between iterations that was within the precision of the computer. Computations were performed using MATLAB® Version 8.2.0.701 (R2013b) [22] running under 64-bit Microsoft Windows 7 Enterprise (SP-1) on a Dell Precision T5500 workstation, with two hexa-core, 2.66 GHz Intel® Xeon processors and 24.0 Gbyte RAM. The NVIDIA Quadro FX 5800 video adapter contained 240 GPU cores, but these were not employed during data processing.

2.3. Results

Three- and four-factor analyses were performed on the influenza types A H1 and H3 data and results are shown in Figures 4-7. This effort is aimed at demonstrating the feasibility of exploring and interpreting data using multi-way modeling. We make no attempt at interpreting the results in terms of epidemiology or inferring biological drivers such as human disease transmission, or virus lifecycle. However, we expect that some of the factor patterns we uncover are related to well-understood biological phenomena.

In all cases, convergence occurred in less than 400 iterations and less than 1.5 seconds. During each iteration, the region and year mode factors were normalized to unit-vector length to prevent issues related to numerical precision. Thus, the magnitude related to each factor was imposed on the week mode vectors. Higher rank analyses were performed for both types, A (H1) and A (H3), but are not presented in this limited report.

Figures 4-7 display models of the influenza data. The top of each figure shows the weekly, temporal, mode for the decomposition. The lower left box displays the regional mode loadings. All markers are the same color as the regions identified in Figure 3 and the connecting lines are the colors of the various factors identified in the legend in the top box. The lower right plot contains the annual mode loadings.

Figure 4 contains the 3-factor NNTF of the influenza type A (H1) data. The following is an interpretation of each factor of this model.

- Factor 2 contains a dominant early rise to the flu season. This factor has a gradual increase in flu occurrences beginning week 43 and then demonstrates a more rapid rise at week 49, eventually reaching an interim peak at week 52 before it rises to its last maximum at week 6. Inspection of this same factor (factor 2) in the regional mode shows that regions 4, 5 and 9 are the locations of this factor 2 behavior. These regions correspond to the north central, southeast, and west coastal areas of the United States. This pattern occurs almost exclusively in years 2001, 2007 and 2008 and to a lesser extent in 2003.

- Factor 1 contains the highest peak of the temporal mode factors. Factor 1 rises sharply from week 51 with a shoulder at week 1 before it goes on to peak at week 6 or early- to mid-February. It then experiences a similarly sloped decline through week 16 before a one week rebound in week 17, or mid-March. The regional mode for factor 1 shows the primary affected regions are 3 and 5, with mildly affected regions 1 and 6. This factor was most dominant in 2009 with a smaller contribution in 2003.

- Factor 3 rises slowly from week 41 to an initial weak shoulder at week 52, it then rises to a peak at week 7. Finally, the flu season ends with a substantial 3-week rebound during weeks 17-19, peaking in week 18. It takes place in regions 4 and 9, with smaller incidence in regions 7, 8 and 10. This factor is dominant in 2009 and to a lesser extent in 2003, similar to factor 1.

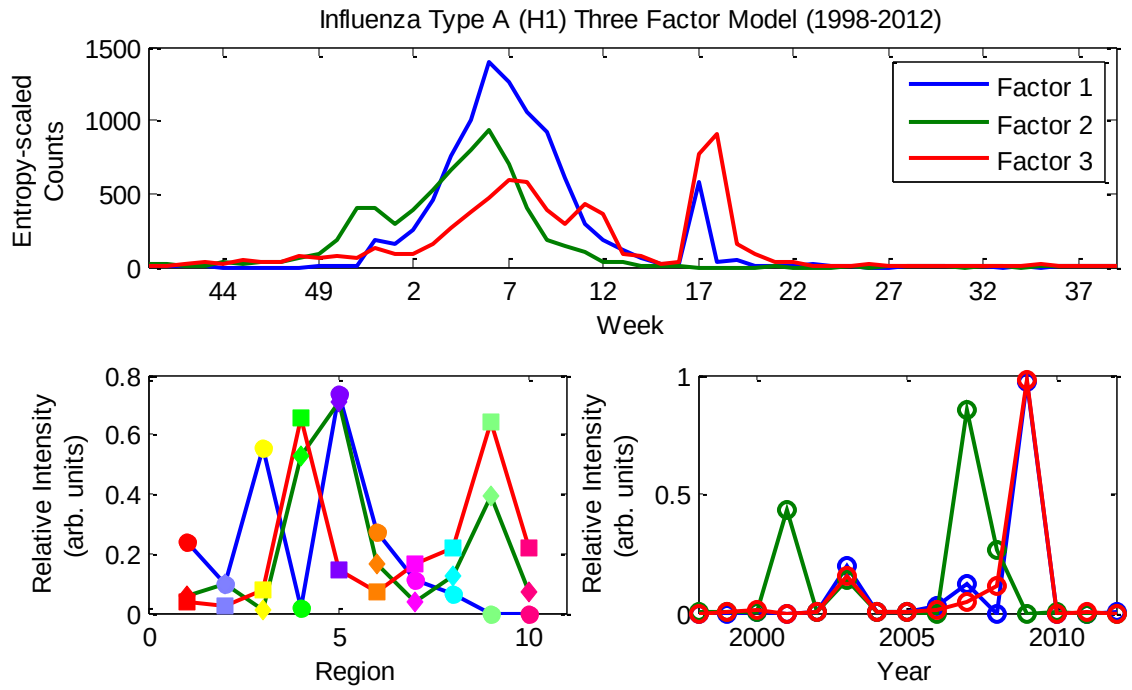

Figure 4. 3-factor NNTF model of influenza type A (H1) for years 1998-2012.

Figure 5 contains the 4-factor NNTF of the influenza type A (H1) data. The factors have been ordered to optimally match the factors in Figure 4.

The following is a description of the 4-factor model behavior (Figure 5),

- Factor 2 is the most stable of the factors with this increase in rank from 3- 4. There is a gradual, early rise to the flu season from week 43 and then a more rapid rise at week 49, eventually reaching an interim peak at weeks 51 and 52 before it the rises to maximum at week 6. Its regional mode events are practically identical to that of the rank-3 model and reside in regions 4, 5 and 9. This pattern occurs almost exclusively in the same years: 2007 and 2008 with a mild prevalence shift from 2001 to 2003.

- Factor 1 still exhibits the largest peak in the temporal mode; however it has lost the shoulder prior to its peak at week 6. It declines through week 16 before the rebound in week 17. The factor 1 regional mode again shows that regions are 3 and 5 are primarily affected, with minor effects in regions 1 and 6. Its annual mode is essentially unchanged.

- The temporal mode of factor 3 sees a similar fate as factor 1: diminishment of the early season shoulder, with retention of the peak and late season rebound. It retains its spatial distribution peaking in regions 4 and 9. It also occurs predominantly in 2009 and to a lesser extent in 2003.

- Factor 4 appears to have arisen from factors 1 and 3 as a distinct factor. In the temporal mode, it exhibits a shoulder at week 52 and later peaks at week 4. This essentially expunges the early shoulders in factors 1 and 3. Later, in week 7, it participates in the late winter rebound. The regions most impacted are 5 and 6, however all regions are involved in this factor. This factor occurs predominantly in 2009 and to a much smaller degree in 2001.

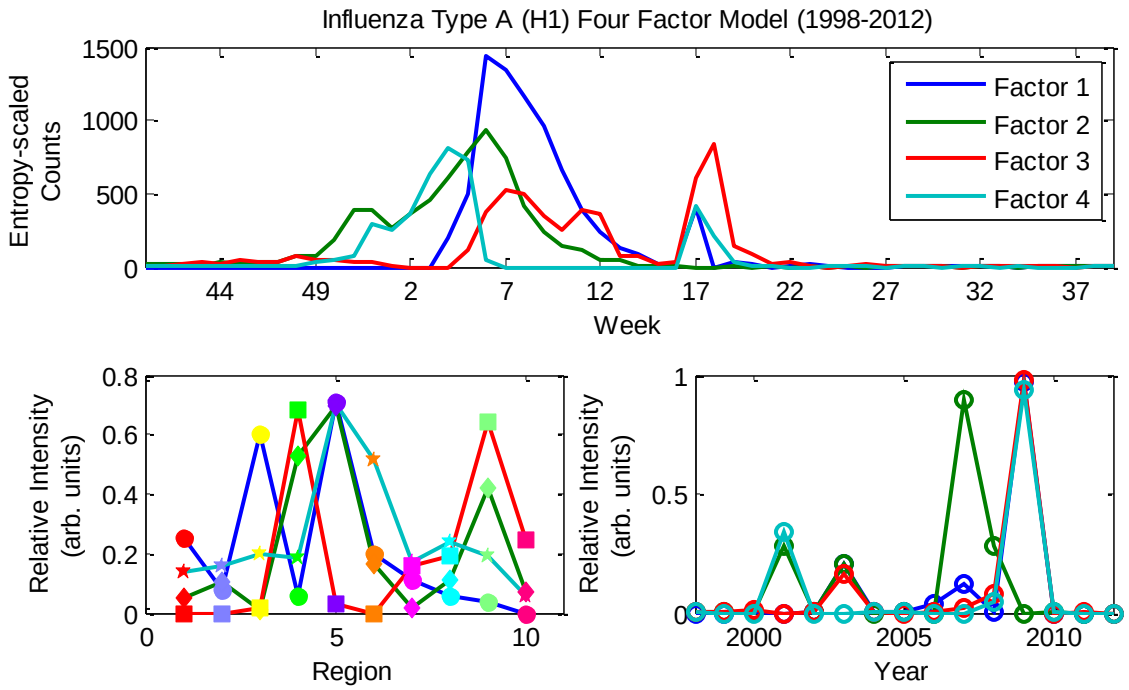

Figure 5. 4-factor NNTF model of influenza type A (H1) for years 1998-2012

Figure 6 represents the 3-factor NNTF of a different data set: the influenza type A (H3) data. The temporal mode displays three fundamentally unimodal peaks.

- Factor 1 is centered near week 51 and has full-width half maximum of approximately 5 weeks. This factor arose primarily in 2004, with minor events in 1998, 2000, 2005, 2006 and 2011. Factor 1 is principally in regions 5 and 4, with a mild presence almost everywhere.
- Factor 2 is centered near week 5 and is about 10 weeks wide. Factor 2 is also present in almost all regions with greatest strength in 3 and 6.
- Factor 3 centers at week 10 with a width of 7 weeks. Factor 3 resides chiefly in region 5, but is present consistently in regions 3, 4, 7, 8, 9 and also weakly in region 10.

Figure 7 displays the 4-factor NNTF of the influenza type A (H3) data. Factors 1, 2 and 3 are practically indistinguishable to those respective factors of the 4-factor model (Figure 6). Indeed, both the region and annual modes are identical in both model ranks three and four. Factor 4 shows a bimodal rebound, independent of the other outbreaks, in the temporal mode. The two peaks of this rebound occur at weeks 18 and 22. This particular rebound emerges primarily in regions 9 and 4, and in years 2009 and with a largely reduced magnitude in 2012.

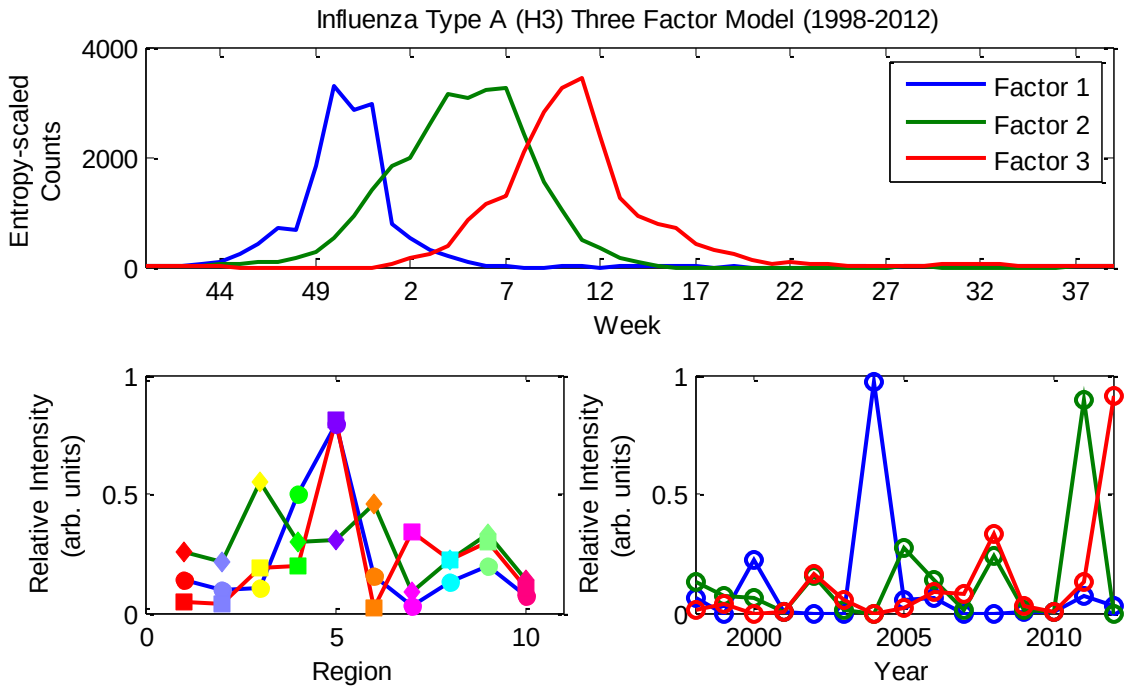

Figure 6. 3-factor NNTF model of influenza type A (H3) for years 1998-2012.

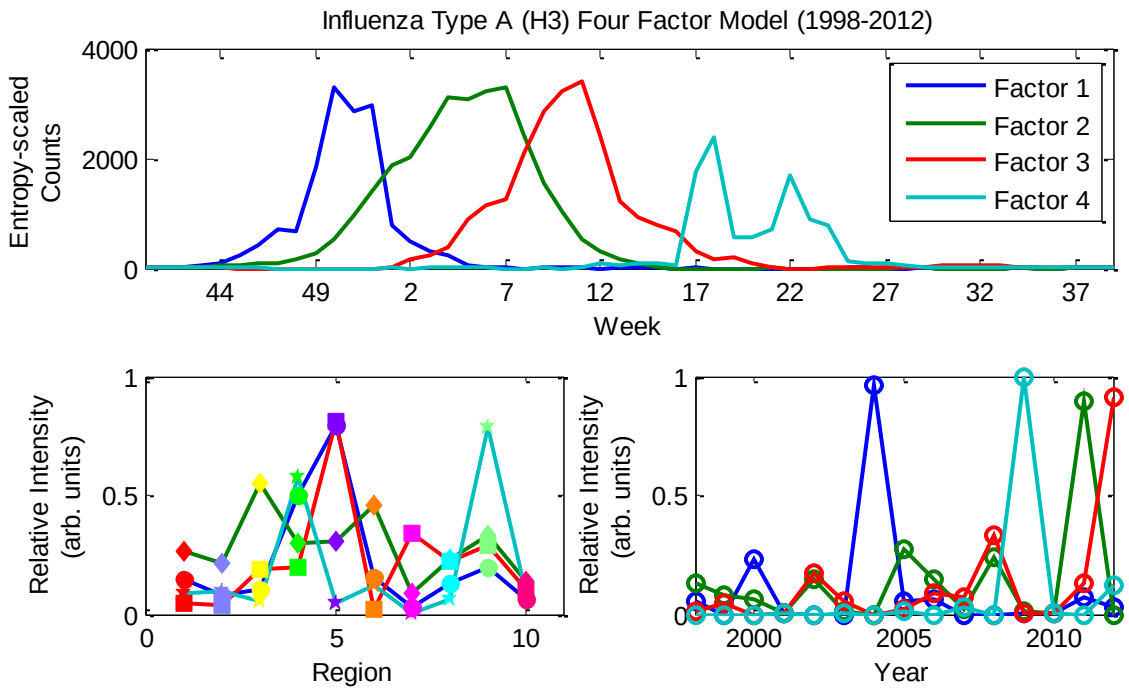

Figure 7. 4-factor NNTF model of influenza type A (H3) for years 1998-2012.

2.4. Discussion

The analyses presented above indicate influenza outbreak patterns are, in general, highly regional. The two influenza types investigated have quite different temporal pattern, which we surmise are related to the

virus lifecycles. The A (H1) strain appears to follow a prolonged multi-mode pattern, in which the outbreak grows slowly at first and appears to stagnate then accelerates to full strength. It also demonstrates a rebound in late winter or early spring as part of the same outbreak. Region 5, in the northern Midwest, is heavily involved in the earliest (factor 2) and largest (factor 1) outbreaks. Regions 4 and 9, the Southeast and Pacific Southwest, respectively, significantly participate in the two early outbreak factors, 2 and 3. Region 3, the Mid-Atlantic States, joins region 5 in the largest infection factor, 1. One interpretation of these patterns, at least in 2003 and to some extent in 2009, is disease introduction in the upper Midwest, Southeast and Southwest, or transference to the Southeast and Southwest from the Midwest. In the warmer southern climes, the outbreaks tend to wane without rebound. In the upper Midwest, however, the disease enters a second stage, one that later experiences a rebound, perhaps due to returns from the South in late winter. Mid-season the infection spreads to the Mid-Atlantic States, which also participate in a second eruption in the early spring.

The A (H3) strain presents a different pattern than A (H1), with rather distinct unimodal waves. Factor 1 is the early season wave that struck regions 4 and 5 in 2004 and to a lesser extent 2000. Factor 2 is the mid-season wave that seems to be in all regions, but heaviest in regions 3 and 6, perhaps after moving on from regions 4 and 5. Factor 3 then is the last season wave that returns to region 5 but is also present in regions 3, 4, 7, 8, and 9. Finally, factor 4 is the spring rebound that struck regions 9 and 4 most severely in 2009.

3. ANALYSIS OF TAXI CAB ACTIVITY TRACES

A second case study was conducted to examine the extensibility of the mathematical methods described above to highly structured transportation network data. Specifically, we wished to examine the capabilities of these methods for site (node) discovery including temporal tagging and for inferring activity patterns in other loosely associated domains in which activity patterns are shaped by human behaviors. A freely downloadable dataset [23] of activity traces from taxi cab activity in San Francisco, CA was chosen as an exemplar dataset. The dataset was made available to the public through the Community Resource for Archiving Wireless Data at Dartmouth (CRAWDAD) [24] by Dartmouth College. The dataset contains GPS coordinates of 536 taxis collected over 30 days in the San Francisco Bay Area. Each taxi was equipped with a GPS receiver and sent a location-update (timestamp, identifier, geo-coordinates, occupancy status, i.e. have fare on board or no fare on board) to a central server. An example of the GPS traces from one taxi over the data collection period is shown in Figure 8.

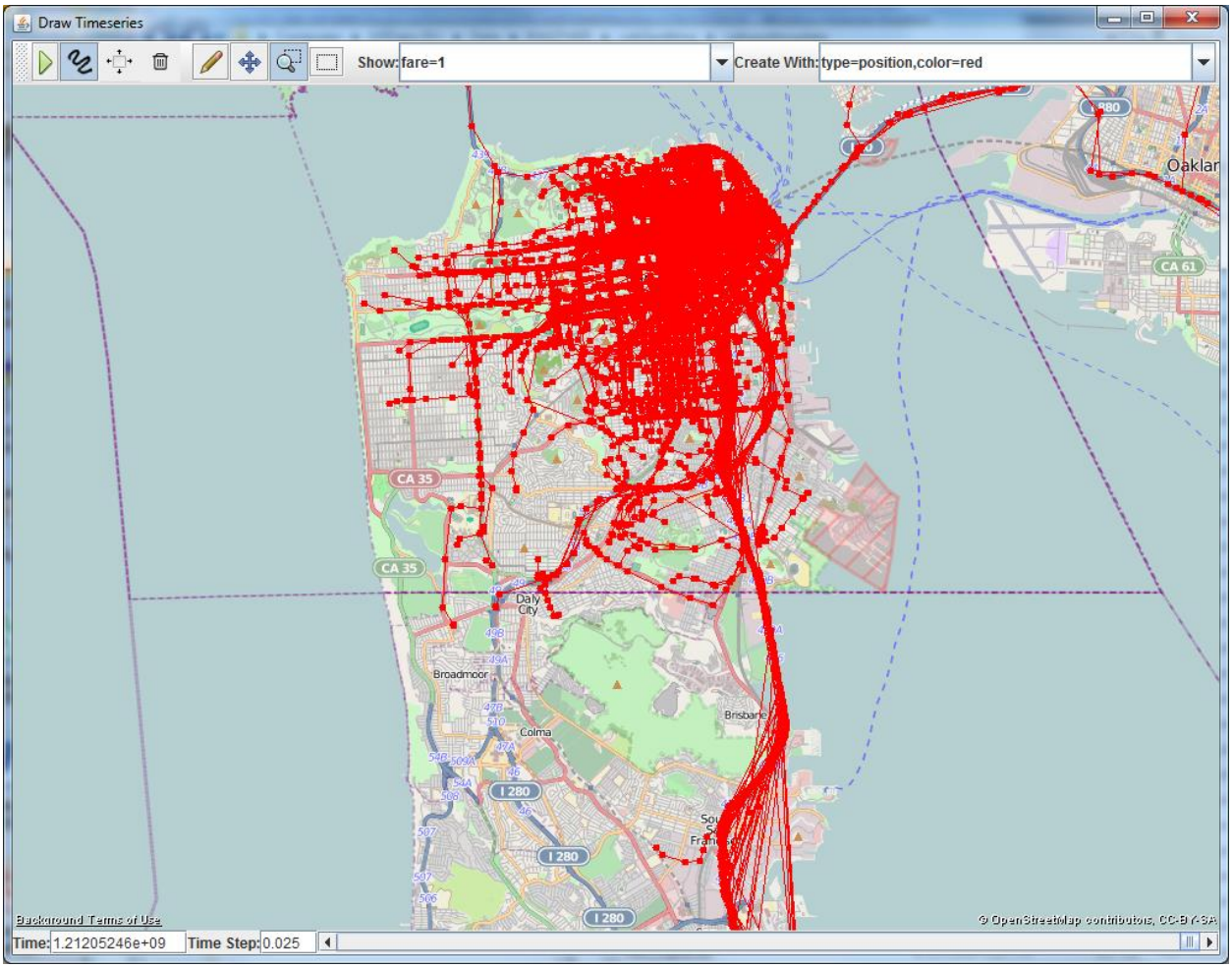

Figure 8. Example GPS traces (in red) from one taxi in the San Francisco Bay Area. The figure was produced using the Relational Blackboard tool (RBB) [25, 26].

3.1. Methodology

As an exemplar problem, we investigated patterns of activity associated with taxi trips departing from the San Francisco International Airport (SFO). For the tensor decomposition, the data where arranged in a 3 dimensional cube (a 3-way or 3rd order tensor) with fare state (0 denoting no fare on board, 1 denoting with fare on board), hour of day and geographical region as the three axes. The geographical region was represented as an index into a geospatial grid with each grid cell encompassing 0.01 degrees longitude by 0.01 degrees latitude. The resulting geospatial grid then contained 7371 cells (81 x 91) over the longitude range from W-122.8 to W -121.9 and latitude range of N37.1 to N38.0. Each element of the tensor contained a count of the number of trips from SFO which ended in the specific geospatial cell, in the specific hour of day with the specific fare state. Figure 9 shows the geospatial distribution of the data tensor where the count data have been scaled by log (base10) to increase the visibility of cells with relatively low trip counts.

3.2. Experimental

The taxi cab trace data were preprocessed as described above organizing them as a fare state by hour of day by geographical region index array with respective dimensions of 2x24x7371. The data were then scaled by adding one and then taking the log base 10 (equation 5).

$$d_{ijk} = log10(d_{ijk} + 1) \tag{5}$$

The preprocessed data were analyzed using non-negative CANDECOMP/PARAFAC (CP) decomposition provided in the MATLAB Tensor Toolbox from Sandia National Laboratories [27]. A 2-factor decomposition was chosen due to the binary nature of the fare state variable. Computations were performed using MATLAB®[22] running under 64-bit Microsoft Windows 7 Enterprise (SP-1) on an HP Z400 workstation, with one quad-core, 2.66 GHz Intel® Xeon processors and 16.0 Gbyte RAM.

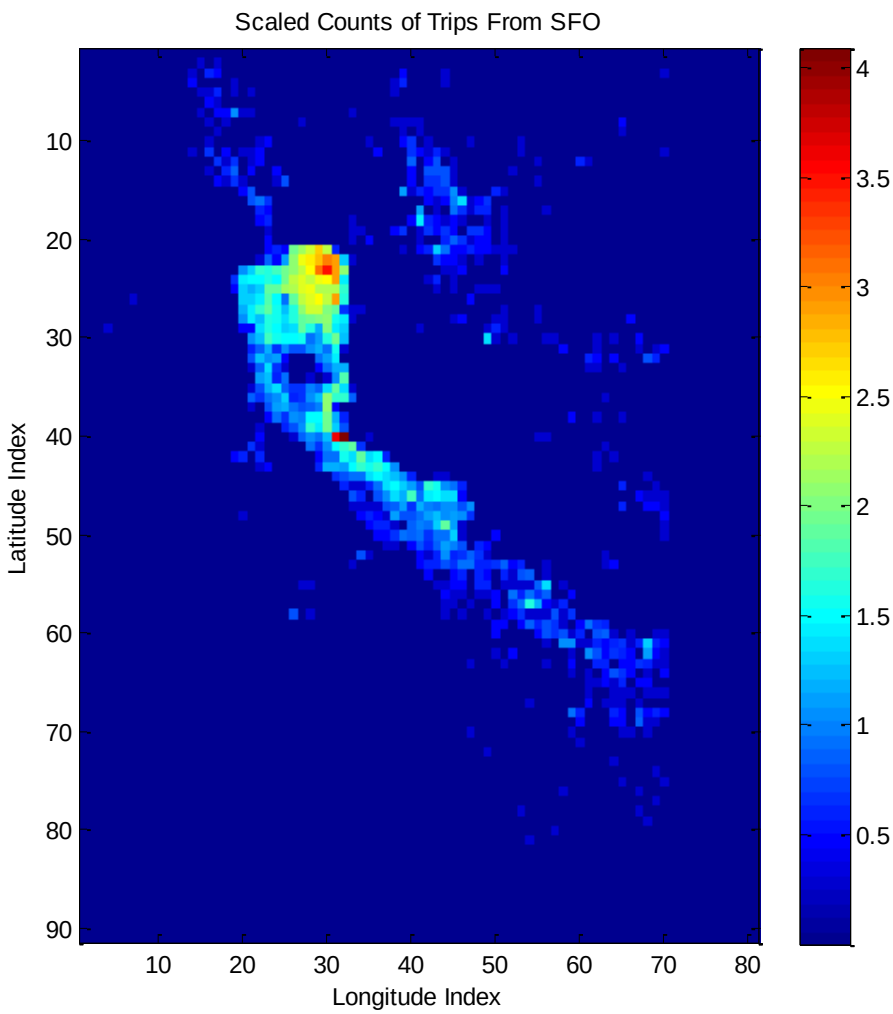

Figure 9. Visualization of geospatial grid input data.

3.3. Results and Discussion

Examination of the resulting factors along each dimension reveals interesting patterns of activity that are obscured in a simple geospatial grid of activity levels. Along the fare dimension, factor 1 describes activity of cabs without fares on board and factor 2 describes activity of cabs with fares on board (Figure 10).

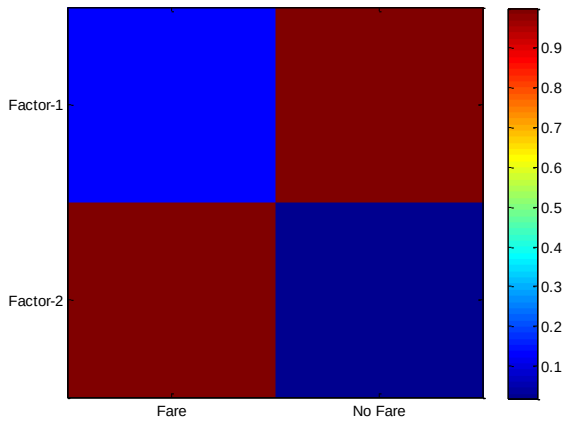

Figure 10. Factors 1 and 2 along the taxi fare dimension.

Along the geospatial dimension, factor 1 shows that nearly all trips from SFO without fares on board terminate within the city of San Francisco boundaries (figure 11). While this may at first seem like a logical result of taxi drivers seeking the most populated areas for new fares, this behavior is also consistent with the participating cab company's license which is restricted to operate in San Francisco and SFO airport only (http://yellowcabsf.com/index.php/faqs). Factor 2 along the geospatial dimension shows the majority of trips from SFO with fares on board terminate in the city of San Francisco (Figure 12), but a number of other important locations outside of the San Francisco city limits are commonly visited.

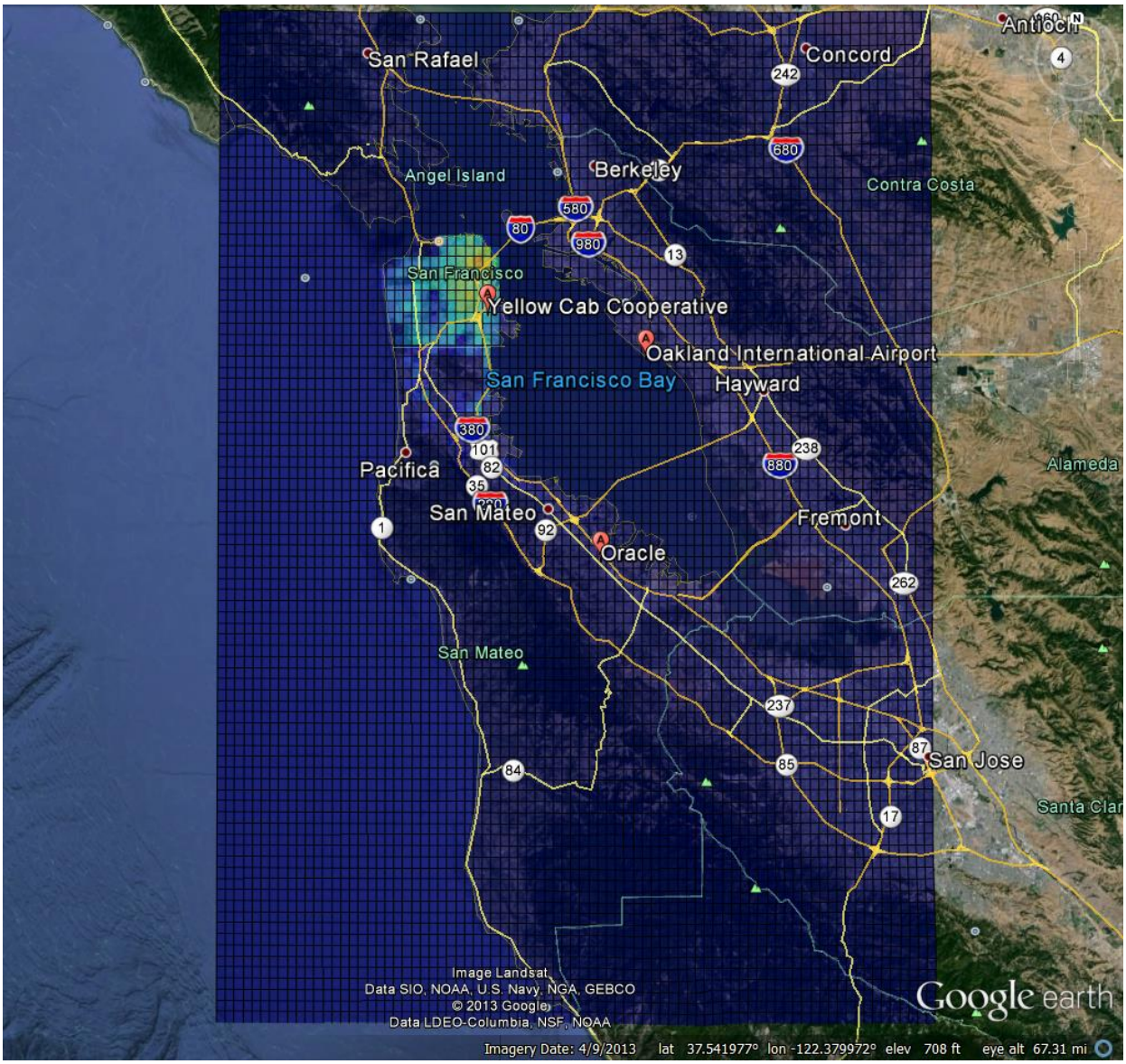

Figure 11. Geospatial patterns of activity described by factor 1 (without fare).

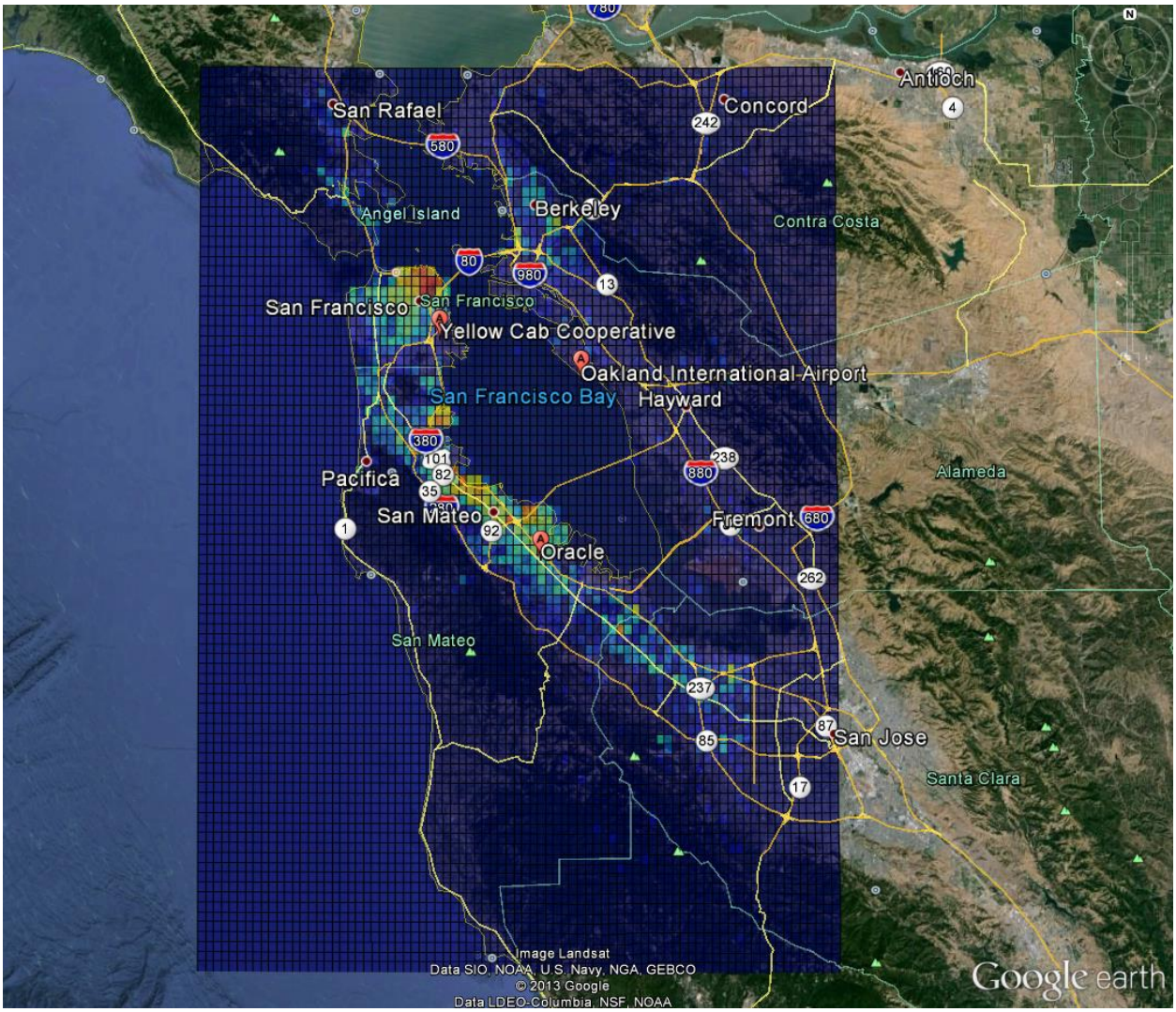

Figure 12. Geospatial patterns of activity described by factor 2 (with fare).

Further investigation of factor 1 along the geospatial dimension provides a clue to which cab company provided the GPS trace data (Figure 13). Zooming in on the geospatial display shows one particular cell receiving a high number of taxi visits when no fare is on board. Using Google Maps, one notices that the Yellow Cab Cooperative business is located in this cell of higher activity. The hypothesis that the Yellow Cab Cooperative provided the cab traces is confirmed in the documentation provided with the dataset download from CRAWDAD.

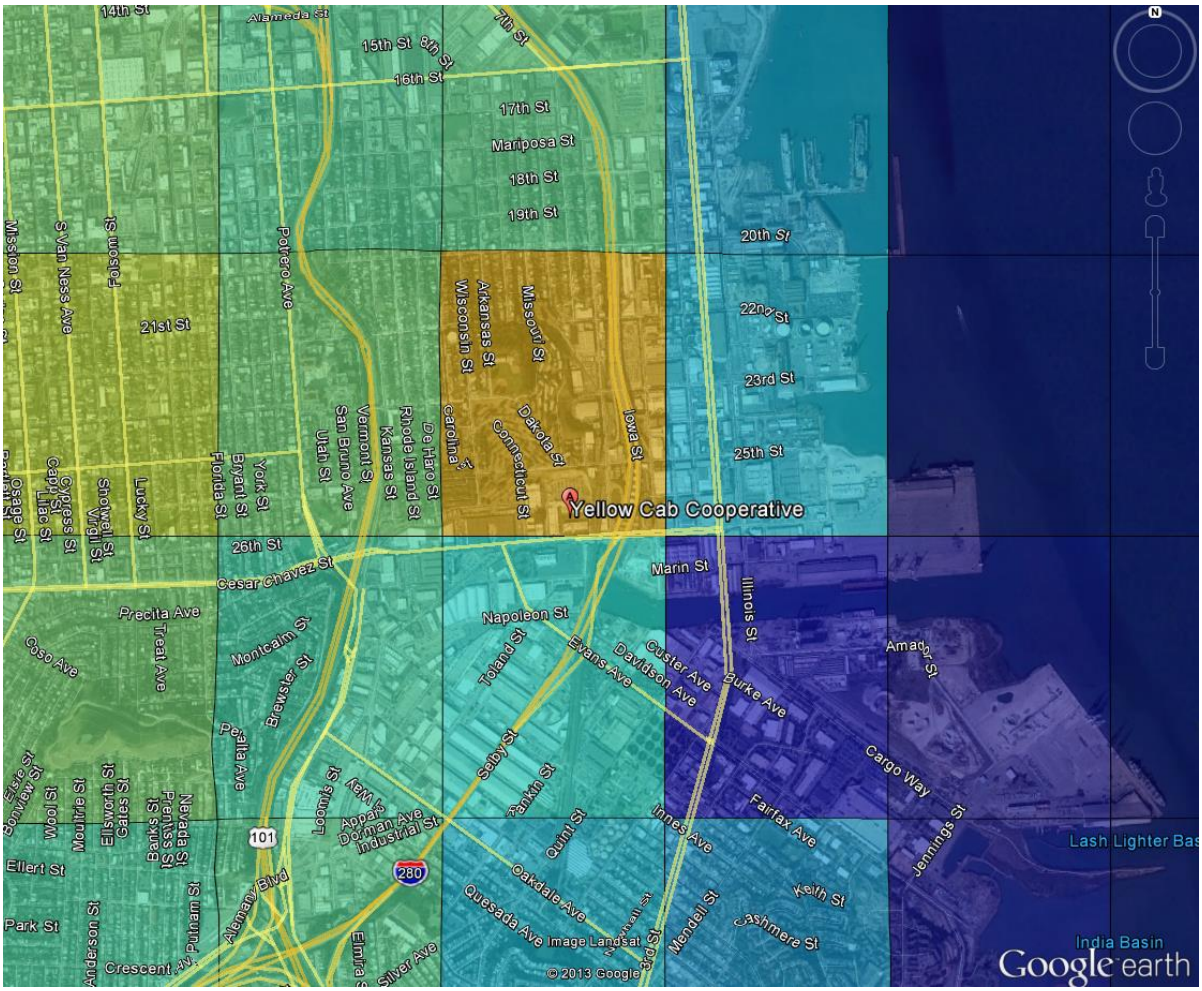

Figure 13. Zoomed view of factor 1 (without fare).

Further investigation of factor 2 along the geospatial dimension reveals a rather high level of taxi trips with fares on board terminating in the vicinity of the Oracle Corporation World Headquarters in Redwood City, CA (Figure 14).

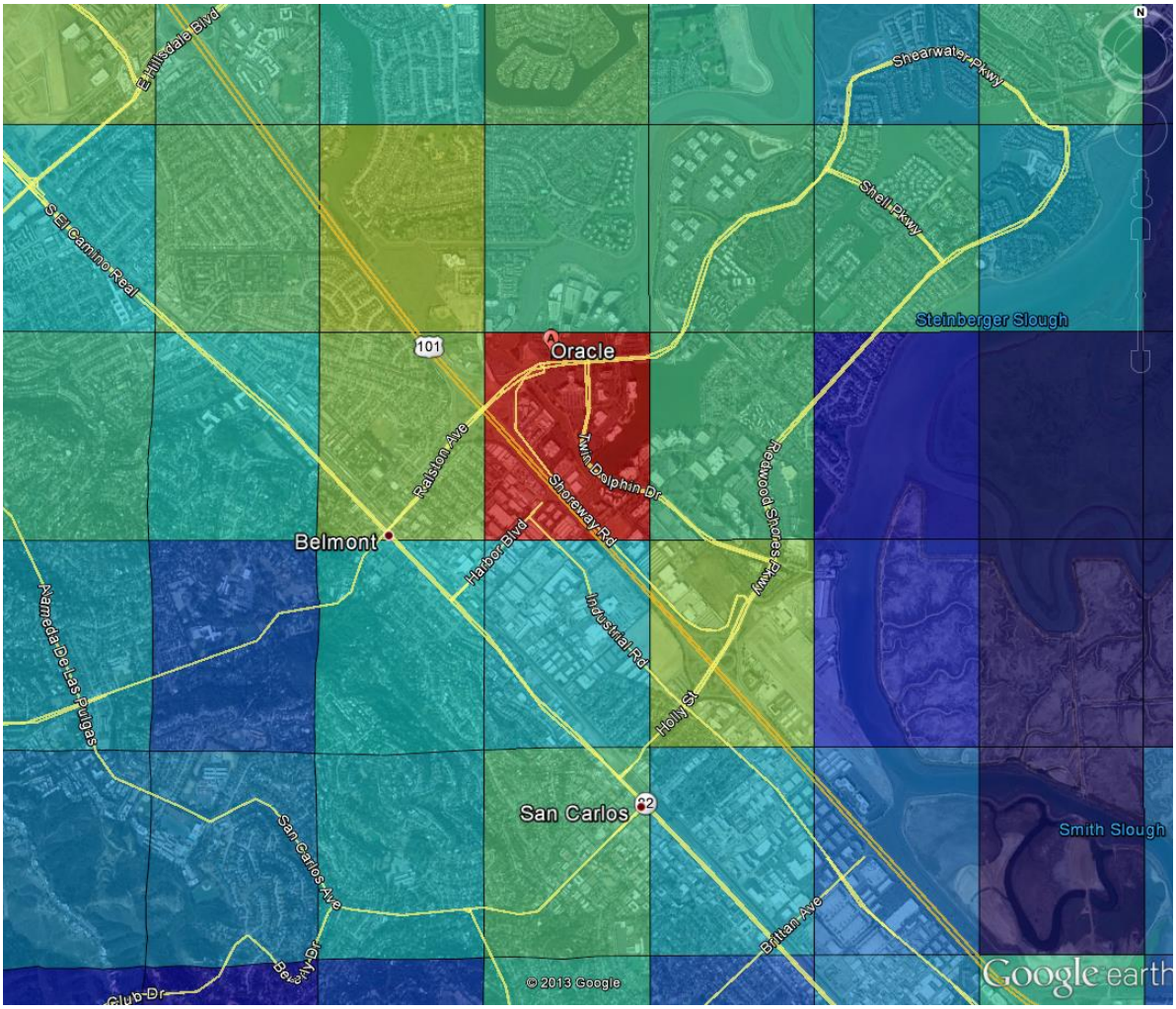

Figure 14. Factor 2 (with fare) along the geospatial dimension.

Factor 2 along the geospatial dimension also reveals a rather unusual level of taxi trips with fares on board terminating directly east, across the San Francisco Bay from SFO (Figure 15). Surprisingly, this active cell contains few destinations other than the Oakland International Airport (OAK), raising the question, "Why would someone take a taxi from SFO to OAK?" We conjecture that this particular behavior may be due to international travelers seeking the lowest travel costs to further destinations once they have arrived in the continental US. One traveler posted on Fodor's Travel Talk forum that they found substantially less expensive rental car rates at OAK (http://www.fodors.com/community/united-states/getting-from-sfo-to-oak.cfm).

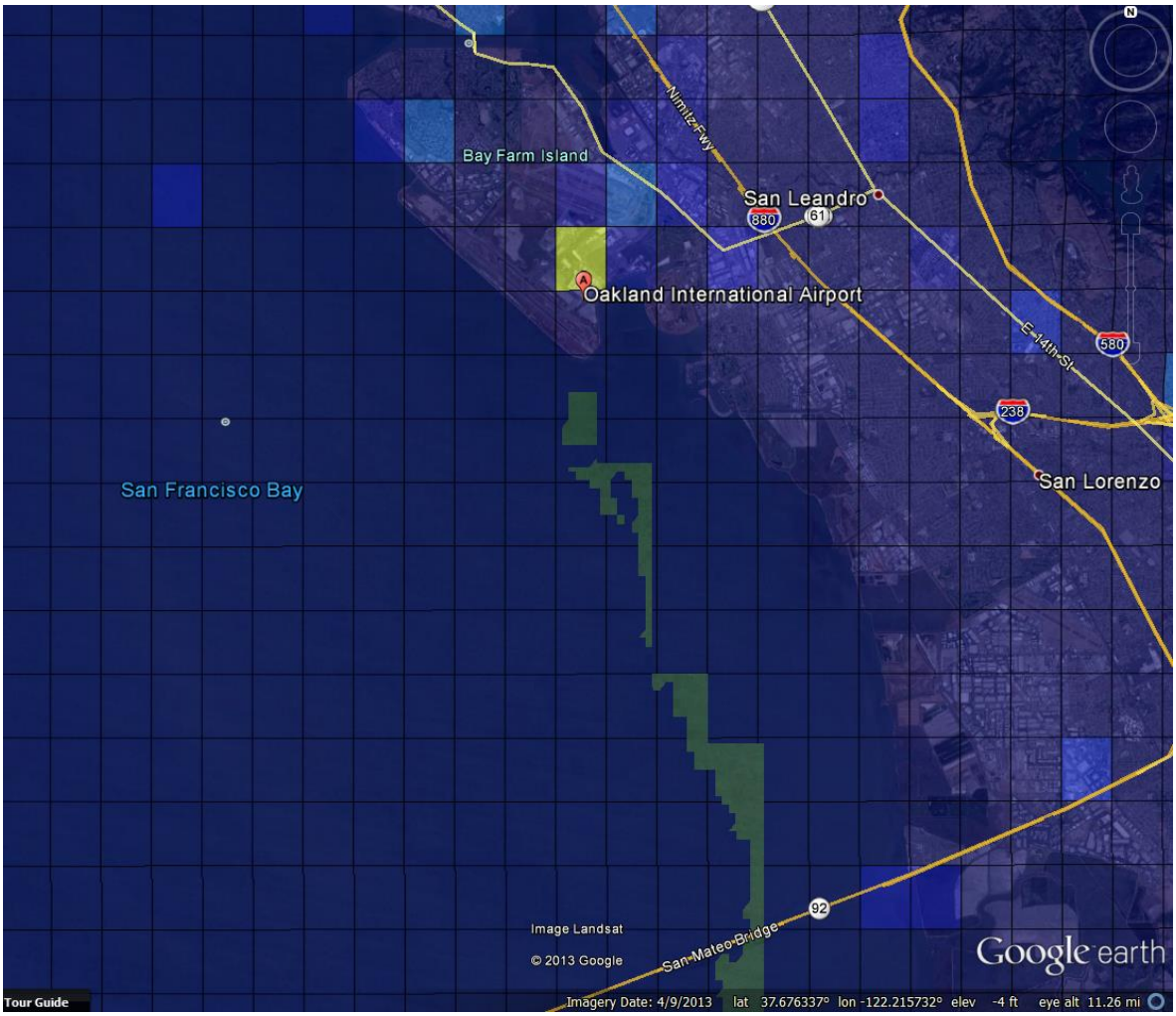

Figure 15. Factor 2 (with fare) along the geospatial dimension highlights a relatively high level of taxi service with fares on board from SFO to OAK.

Along the temporal dimension, Figure 16 shows the relative magnitude of activity of trips from SFO at each hour of the day. Factor 1 (top row) shows trips from SFO without fares on board while factor 2 (bottom row) shows trips from SFO with fares on board. The strong dip in activity in factor 2 between the hours of 5 to 9 AM are consistent with most travelers embarking on departing flights from SFO in the morning hours while relatively few inbound flights arrive during this same time period. Because there are fewer inbound flights than outbound flights at this time of day, taxi drivers arriving at SFO are compelled to return to the city limits without a fare on board.

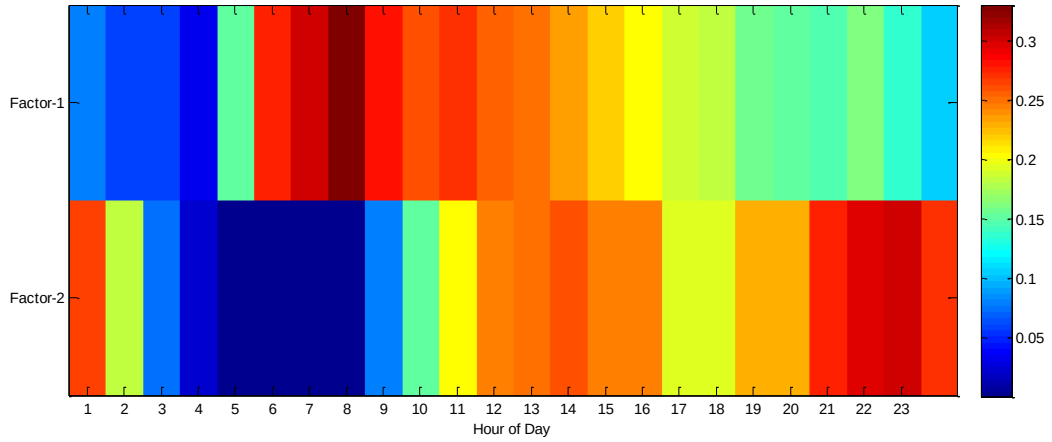

Figure 16. Factors 1 (without fare) and 2 (with fare) along the temporal dimension.

Figure 17 provides another view of the temporal activity of taxi trips departing SFO. For this case, the temporal dimension of the tensor was constructed using the day of the week to accumulate counts of taxi trips from SFO. The resulting tensor size remained the same for the fare and geospatial dimensions, but the temporal dimension was reduced to seven elements (one for each day of the week). Examination of factor 2 shows a marked decrease in taxi trips from SFO with fares on board on Fridays and Saturdays.

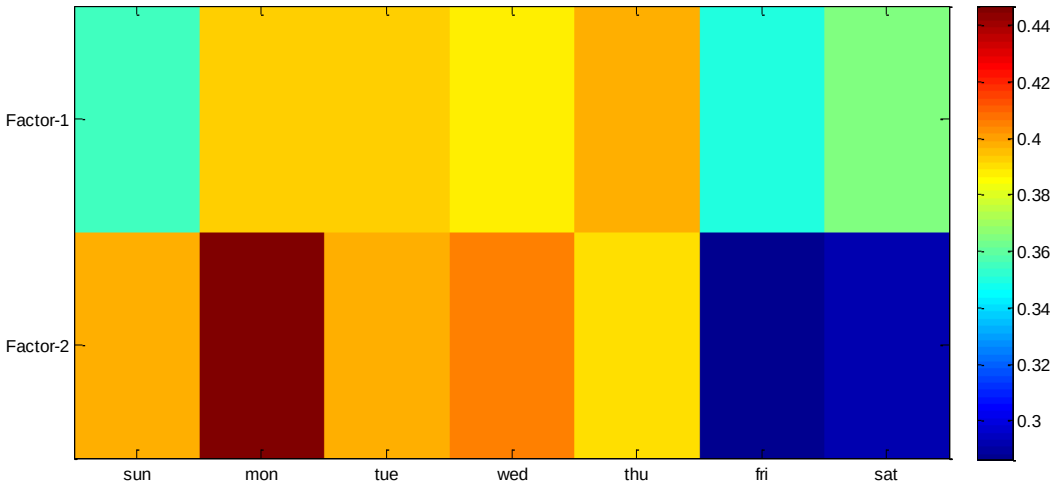

Figure 17. Factors 1 (without fare) and 2 (with fare) along the temporal dimension when trip termination times are binned by day of week.

4. CONCLUSIONS

We have presented results of two case studies where tensor (multiway array) factorization and decomposition techniques have been applied to spatio-temporal datasets. The first case study used influenza data collected over a multi-year period within the United States and its territories. This initial, limited study demonstrates the ability to analyze and visualize the effects of communicable disease outbreak in terms of spatio-temporal progression. Future work should involve larger samples and disease types as well as more in-depth analyses of data collection constraints and the underlying drivers of virus transmission, as documented in the spatial epidemiology literature.

The second case study used taxi cab trace data collected over multiple weeks in the San Francisco Bay Area. By examining taxi trips originating from SFO, this initial study demonstrates the ability to identify patterns of activity related not only to the taxis themselves but also to air passenger traffic. In addition to overall spatial and temporal patterns of activity, these methods have demonstrated an ability to gain insights into other information that shaped the observed patterns of activity, including which taxi cab company participated in the data collection and the limitations placed on this company's area of operations due to municipal licensure restrictions.

Further work should:

- Explore different approaches to defining the geospatial grid used to construct the data tensor,
- Seek to understand how the definition of the geospatial grid may influence the types of activity patterns that may be discovered using tensor analysis techniques,
- Test the tensor model approaches on a portion of the data and determine if the factors can predict taxi or virus behavior in the remaining data and
- Seek to understand the uncertainty in the tensor approach for these datasets (and others)

5. REFERENCES

1. R. A. Harshman and M. E. Lundy, "Parafac : Parallel Factor-Analysis," *Comput. Stat. Data Anal.* 18(1), 39-72 (1994).
2. R. Bro, "PARAFAC. Tutorial and applications," *Chemom. Intell. Lab. Syst.* 38(2), 149-171 (1997).
3. J. B. Kruskal, "Rank, Decomposition, and Uniqueness for 3-way and N-way Arrays," *in Multiway Data Analysis* R. Coppi and S. Bolasco, Eds., pp. 7-18, North-Holland, Amsterdam (1989).
4. B. C. Mitchell and D. S. Burdick, "Slowly Converging PARAFAC sequences: Swamps and two-factor degeneracies," *J. Chemom.* 8, 155-168 (1994).
5. S. E. Leurgans and R. T. Ross, "Multilinear models: Applications in spectroscopy," *Stat. Sci.* **7**(3), 289-319 (1992).
6. G. Tomasi and R. Bro, "A comparison of algorithms for fitting the PARAFAC model," *Comput. Stat. Data Anal.* 50(7), 1700-1734 (2006).
7. E. Sanchez and B. R. Kowalski, "Tensorial resolution: A direct trilinear decomposition," *J. Chemom*. 4(29-45 (1990)
8. S. E. Leurgans, R. T. Ross and R. B. Abel, "A decomposition for three-way arrays," *SIAM J. Matrix Anal. Appl.* 14(4), 1064-1083 (1993)
9. R. A. Harshman, "Foundations of the PARAFAC procedure: Model and conditions for an ‘explanatory’ multi-mode factor analysis," *UCLA Working Papers Phonetics* 16(1–84 (1970).
10. R. Bro and N. D. Sidiropoulos, "Least squares algorithms under unimodality and non-negativity constraints," *J. Chemom.* 12(4), 223-247 (1998).
11. T.G. Kolda., and B. W. Bader, *Tensor Decompositions and Applications*, SIAM Review 51(3), 455-500, September 2009.
12. D. M. Rao, A. Chernyakhovsky and V. Rao, "Modeling and analysis of global epidemiology of avian influenza," *Environmental Modelling & Software* 24(1), 124-134 (2009).
13. C. E. Mills, J. M. Robins and M. Lipsitch, "Transmissibility of 1918 pandemic influenza," *Nature* 432(7019), 904-906 (2004).
14. S. Dorjee, *et al.*, "A Review of Simulation Modelling Approaches Used for the Spread of Zoonotic Influenza Viruses in Animal and Human Populations," *Zoonoses and Public Health* 60(6), 383-411 (2013).
15. F. Brauer, "Epidemic Models with Heterogeneous Mixing and Treatment," *Bull. Math. Biol.* 70(7), 1869-1885 (2008).
16. W. O. Kermack and A. G. McKendrick, "A Contribution to the Mathematical Theory of Epidemics," *Proceedings of the Royal Society of London. Series A* 115(772), 700-721 (1927).
17. "CDC FluView Weekly Report," Centers for Disease Control and Prevention, Ed., CDC, Atlanta, GA (2012).
18. C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal* 27(3), 379-423 (1948).
19. E. T. Jaynes, "Information Theory and Statistical Mechanics," *Physical Review* 106(4), 620-630 (1957).
20. M. H. Van Benthem and M. R. Keenan, "Tucker1 model algorithms for fast solutions to large PARAFAC problems," *J. Chemom.* 22(5), 345-354 (2008).
21. M. H. Van Benthem and M. R. Keenan, "Fast algorithm for the solution of large-scale non-negativity-constrained least squares problems," *J. Chemom.* 18(10), 441-450 (2004)
22. "MATLAB," The MathWorks, Inc., Natick, MA (2013).
23. M. Piorkowski,N. Sarafijanovic-Djukic, M. Grossglauser, *A Parsimonious Model of Mobile Partitioned Networks with Clustering*, The First International Conference on COMmunication Systems and NETworkS (COMSNETS), Bangalore, India, 2009.
24. http://crawdad.cs.dartmouth.edu/index.html

25. Relational Black Board software (RBB), Copyright 2012 Sandia Corporation. Under the terms of Contract DE-AC04-94AL85000 with Sandia Corporation, the U.S. Government retains certain rights in this software.
26. R.G. Abbott. "The Relational Blackboard." BRIMS 2013, July 2013. Carlton University, Ottawa, Canada.
27. B. W. Bader, T. G. Kolda and others. *MATLAB Tensor Toolbox Version 2.5*, Available online, January 2012. URL: http://www.sandia.gov/~tgkolda/TensorToolbox/index-2.5.html.

DISTRIBUTION

1	MS0359	D. Chavez, LDRD Office	1911
1	MS0899	Technical Library	9536 (electronic copy)
1	MS0519	S. Castillo	5346
1	MS0519	K. Czuchlewski	5346
1	MS0519	J. Chow	5349
1	MS0519	L. McNamara	5346
1	MS0576	J. Feddema	5511
1	MS1244	B. Malm	5572
1	MS1326	R. Brost	1464
1	MS1327	D. Stracuzzi	1462
1	MS1327	P. Bennett	1463
1	MS1327	B. Hart	1464
1	MS1411	A. Sun	1814

Sandia National Laboratories