Fine-scale visualization of pollen concentrations across the Eastern United States: A space-time parallel approach

M.R. Desjardins[1], A. Hohl[1], A. Griffith[1], and E. Delmelle[1]*

[1]Center for Applied Geographic Information Science, Department of Geography and Earth Sciences, University of North Carolina at Charlotte, 9201 University City Blvd., Charlotte, NC 28223, United States
*Email: eric.delmelle@uncc.edu

Abstract

Allergic rhinitis (hay fever) resulting from seasonal pollen affects 15-30% of the population in the United States, and can exacerbate several related conditions including asthma, atopic eczema, and allergic conjunctivitis. In this paper, we reconstruct the dynamics of pollen concentrations across the Eastern United States at a very fine scale by interpolating daily pollen counts in space and time, obtained from a custom web scraper from February 3rd, 2016 toDecember 14th, 2016. We conducted a space-time cross-correlation and inferred the optimal spatial and temporal range at which correlation vanishes. Given the sheer volume of the computation requirement, we adopt a parallel computational approach facilitated by a spatiotemporal domain decomposition algorithm. We visualize the results in a 3D-enviornment, revealing the space-time patterns of pollen season. This method improves the understanding of large-scale seasonal pollen patterns that may aid physicians with treatment plans for sensitive patients, such as limiting outdoor exposure or physical activity. Our approach is portable to analyze other large spatiotemporal explicit datasets obtained from the web, such as air pollution and precipitation.

Keywords: GIS, space-time interpolation, pollen, parallel computing, 3D visualization

1. Introduction

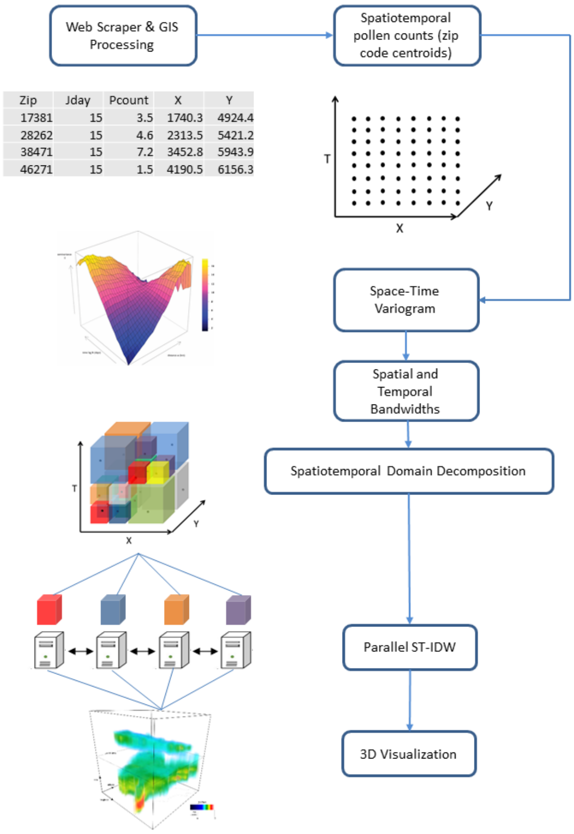
Allergic rhinitis (hay fever) resulting from seasonal pollen affects 15-30% of the population in the United States (US), and causes or exacerbates several associated conditions including asthma, atopic eczema, and allergic conjunctivitis (Wheatley and Togias 2015). Allergic rhinitis is responsible for approximately 2 million missed school days and 3.5 million missed days of work in the US, annually (Nathan, 2007). Allergy to pollen is contained within the broader respiratory health issue and works in conjunction with asthma, air pollution, and chronic obstructive pulmonary disease weaving a complex fabric of breathing-related health risks (Charpin and Caillaud, 2014). Accurately monitoring and predicting pollen counts can aid physicians to develop treatment plans for their patients and inform allergy sufferers to limit physical activity and outdoor exposure (Levetin and Van de Water 2003).

Pollen samples are collected at monitoring stations by an air sampling device. The pollen concentration is determined by counting the number of grains per cubic meter of air, and subsequently, the concentration is converted to an index. Current pollen detection stations are sparsely distributed. In the United States, the National Allergy Bureau collects data at various locations but not every state has a pollen monitoring station. It is, therefore, essential to predict pollen concentrations at unmonitored times and locations, especially since pollen distribution is a continuous phenomenon.

Interpolation techniques can estimate pollen concentrations at unsampled locations based on the values at known locations (Goovaerts 1997; Kyriakidis and Journel 1999). In the interpolation process, the importance of locations where pollen is measured is directly a function of (1) the distance to the location where pollen is being estimated (2) the temporal difference between the sampled and unsampled locations. Spatiotemporal interpolation has been used extensively to study meteorological processes (Hussain et al. 2010; Cao et al.

2015) and disease (Gething et al. 2007), air pollution and air quality such as concentrations of PM 2.5 (Li et al. 2014), ozone (Fang and Lu 2011), nitrogen dioxide (Pebesma et al. 2007) and carbon dioxide concentrations (Guo et al. 2015). However, research regarding spatiotemporal interpolation of pollen concentrations is scarce (Alba et al. 2006; Garcia-Mozo et al. 2006; Siska et al. 2006; DellaValle et al. 2012; León Ruiz et al. 2012; Siska et al. 2012; Aguilera et al. 2015; Rojo and Pérez-Badia 2015). For example, DellaValle et al. (2012) interpolated pollen counts from 14 allergy monitoring stations using ordinary kriging in the northeastern and mid-Atlantic region of the United States. Alba et al. (2006) and Rojo and Pérez-Badia (2015) also used kriging approaches to analyze pollen concentrations from Olive plants in Spain.

Spatial analytical methods such as interpolation and clustering techniques can be computationally prohibitive. This may result in unacceptably slow applications that have execution times which explode with increasing resolution and scale of analysis (Hohl et al., 2016). Therefore, analyzing datasets of increasing size, diversity and availability necessitates accelerated processing capabilities, which are offered by high-performance parallel computing (HPC). HPC provides answers to complex and voluminous computational problems within short time, and enables us to extract spatiotemporal patterns of pollen concentration.


Web Scraper & GIS Processing
Spatiotemporal pollen counts (zip code centroids)
T
X
Y
Space-Time Variogram
Spatial and Temporal Bandwidths
Spatiotemporal Domain Decomposition
T
X
Y
Parallel ST-IDW
3D Visualization

Zip	Jday	Pcount	X	Y
17381	15	3.5	1740.3	4924.4
28262	15	4.6	2313.5	5421.2
38471	15	7.2	3452.8	5943.9
46271	15	1.5	4190.5	6156.3

Figure 1: Framework of parallel spatiotemporal interpolation of pollen counts.

We propose a framework (Figure 1) that extracts daily pollen count data from a popular weather website, applies parallel computing techniques to estimate the space-time variation of pollen concentration, and visualize pollen dynamics in a 3D environment to facilitate the identification of distinct allergy seasons. Our framework is flexible and portable to other case studies attempting to estimate the space-time variation of air pollution phenomena. The parallel computing approach developed in this paper is particularly attractive when such estimations must be conducted and at fine scale.

2. Data and Methods

Our study area (Figure 2) is located in the eastern United States across 31 states and the District of Columbia (DC). The eastern United States combines high levels of net primary productivity with high population densities. The set of zip codes was selected following a stratified random sampling approach, and augmented with zip codes in densely populated areas, forming 3,193 zip codes for which daily pollen counts were collected.

Pollen Counts

We collected daily pollen counts from February 3rd 2016-December 14th 2016 (315 days). The data was extracted from Weather Underground’s website using a custom web scraper. The resulting dataset (n=708,562 observations) includes the zip code, pollen count, coordinates of the zip code centroids, and Julian day. The pollen index ranges between zero and twelve, which combines the pollen levels from a variety of pollinating allergenic plant types (e.g. trees, grasses, and weeds) into a single value.

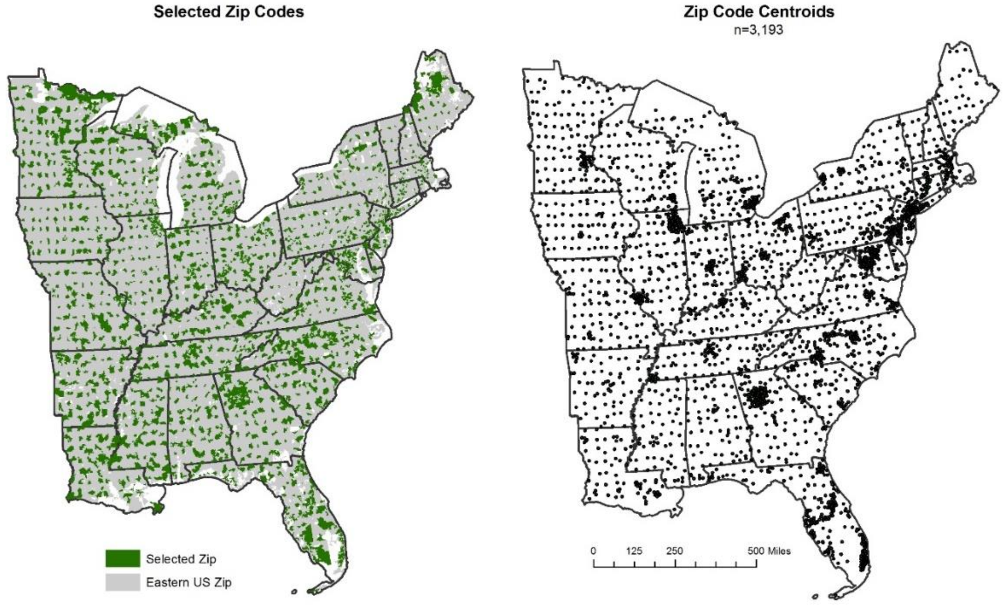


Figure 2: Selected zip codes and their associated centroids (n=3,193). Each centroid acts as a sample point (known location) for the spatiotemporal interpolation.

Space-time Interpolation

To estimate pollen concentrations at unmonitored times and locations, we propose a variation of Li et al.’s (2014) space-time inverse distance weighted approach (ST-IDW):

$$w_r(x,y,t) = \sum_{i=1}^{N_r} \lambda_i w_i \tag{1}$$

$$\lambda_i = \frac{(\frac{1}{\widehat{d_i}})^p}{\sum_{k=1}^{N} (\frac{1}{\widehat{d_k}})^p} \tag{2}$$

$$\widehat{d_i} = \sqrt{(\widehat{x_i} - \hat{x})^2 + (\widehat{y_i} - \hat{y})^2 + (\widehat{t_i} - \hat{t})^2} \tag{3}$$

Equation (1) computes the interpolated value *w* at a three-dimensional location *(x, y, t)*, using pollen count observations at known locations that fall within the specified spatiotemporal bandwidth N_R. λ_i are the weights assigned to the known locations w_i. Note that the known locations w_i are the centroids of the 3,193 zip codes at day *t* shown in Figure 2. In Equation (2), the exponent *p* influences the weight of each known location in the spatiotemporal search radius, which is determined by a space-time variogram. Larger exponents will allocate less weight to points that are farther away from the unknown location. For this study, *p* is assigned a value of 2. Equation (3) calculates the spatiotemporal distance, where both the temporal and spatial distances are normalized between 0 and 1; normalization is necessary because of the very large range of spatial distances (meters) and very small range of temporal distances (days). We apply Equation (3) to a space-time grid of 5km by 5km spatial resolution and 1-day temporal resolution. We compute a space-time variogram to identify the space-time range at which pollen observations do not correlate with one another anymore, essentially calibrating the spatiotemporal bandwidth N_R (Sherman 2011):

$$\hat{\gamma}(h,u) = \frac{1}{2n(h,u)} \sum_{n(h,u)} \{w_i - w_j\}^2 \qquad (4)$$

Given a particular combination of spatial (*h*) and temporal (*u*) separations among observed locations, Equation (4) sums the difference in pollen concentration, and divides that value by *2n,* which the number of pairs of observations within those distance constraints. The space-time variogram was estimated in *R* (using libraries *sp* and *gstat*).

Space-time Domain Decomposition

We split the large and complex task of computing ST-IDW for the pollen dataset into subtasks, which we distribute among multiple processors (CPUs) in parallel. To balance the computational workload among processors and therefore, increase efficiency while decreasing execution time, we perform spatiotemporal domain decomposition. We apply a recursive quadtree decomposition algorithm on the spatial domain of the pollen dataset, while splitting the temporal domain in a regular fashion (Hohl et al. 2016). The spatial distribution of pollen counts is heterogeneous (small zip code areas form clusters in urban areas) while the temporal distribution is not (daily values), except some periods where data collection had stopped. Therefore, it made sense to implement a spatially adaptive and temporally static decomposition Ding and Densham 1996) algorithm.

We choose two parameters: 1. A decomposition threshold (t_d) which should be much lower than the number of pollen counts, 2. A buffer ratio threshold (t_b), which is less than 1. Together, they guide the granularity of decomposition: Low thresholds result in fine-grained decomposition, which facilitates load balancing, as distributing many small tasks will likely result in equal shares among processors than distributing larger tasks. However, this leads to deep levels of recursion and ultimately, a crash of the program. Therefore, we chose t_d = 5000 and t_b = 0.025, which balances the granularity of decomposition versus the danger of stack overflow.

Parallel computing

We use Python and R for software implementation and deploy them on a high-performance computing cluster which has 32 nodes that are connected by an infiniband network switch. Each computing node has 12 CPUs and 12 GBs of memory, resulting in a total of 384 CPUs (Intel Xeon processors, 2.67 GHz clock speed).

3D Visualization
We visualize the results of the interpolation in a space-time cube (Nakaya 2013) to reveal the spatiotemporal patterns of allergy season. Each individual voxel contains a pollen count that was estimated according to Equation (3). Volumetric data are challenging to visualize, however recent advances have proven particularly useful to portray patterns of 3D geographic data (Demsar and Virrantus 2010, Delmelle et al. 2014). Once pollen counts are estimated for each voxel, they can be visualized by color-coding each voxel based on its pollen value. We create a volume of pollen values using a rainbow color scheme. Voxels with a pollen value of 9 and over are colored using dark red, values lower than 5 are colored using dark blue shades, and so on. Given the number of voxels to visualize, we use volume rendering. From a public health perspective, we are interested to visualize regions where pollen counts are particularly elevated. As such, the transparency level of each voxel is adjusted based on its pollen value (voxels with lower interpolated pollen values are assigned a higher level of transparency, whereas higher pollen values are kept opaque).

3. Results

Figure 3 illustrates the result of the space-time variogram that estimates the optimal spatiotemporal bandwidth for our interpolation. We implemented 25 temporal lags of 3 days each and 30 distance lags of 100km each. The temporal component of the variogram flattens out around 45 days and the spatial component flattens our around 1,750km.

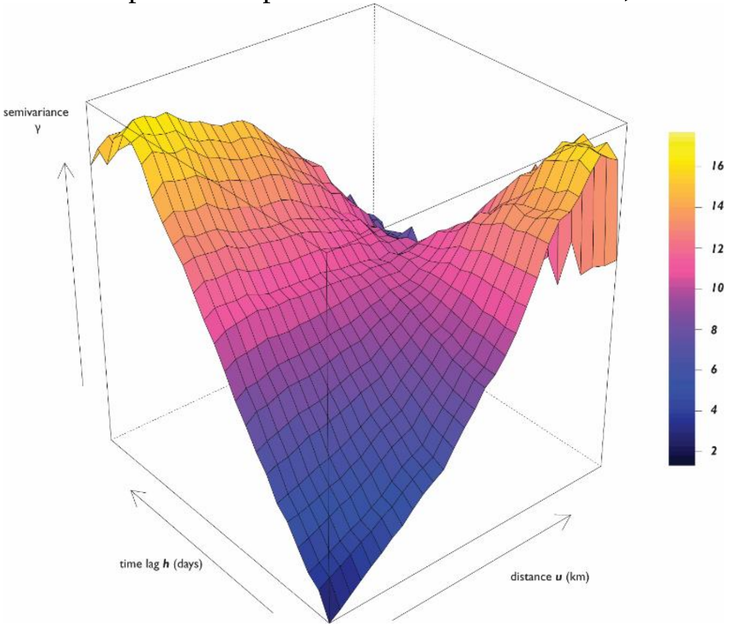


Figure 3: Space-time variogram of pollen data.

Figure 4 provides a 3D visualization of the spatiotemporal interpolation of pollen concentrations, illustrates the continuous dynamics of the pollen season. We observe a cluster of high pollen counts in Florida from early-to-mid spring. There are also smaller clusters of high pollen counts in the northeastern U.S. that occurred in mid-spring to early-summer. This corresponds to the northern U.S. having shorter growing seasons, therefore, the number of days between pollen clusters is lower in the North and higher in the South. Allergy season also occurs later in the northern United States because of a variety of climatic factors (e.g. temperature).

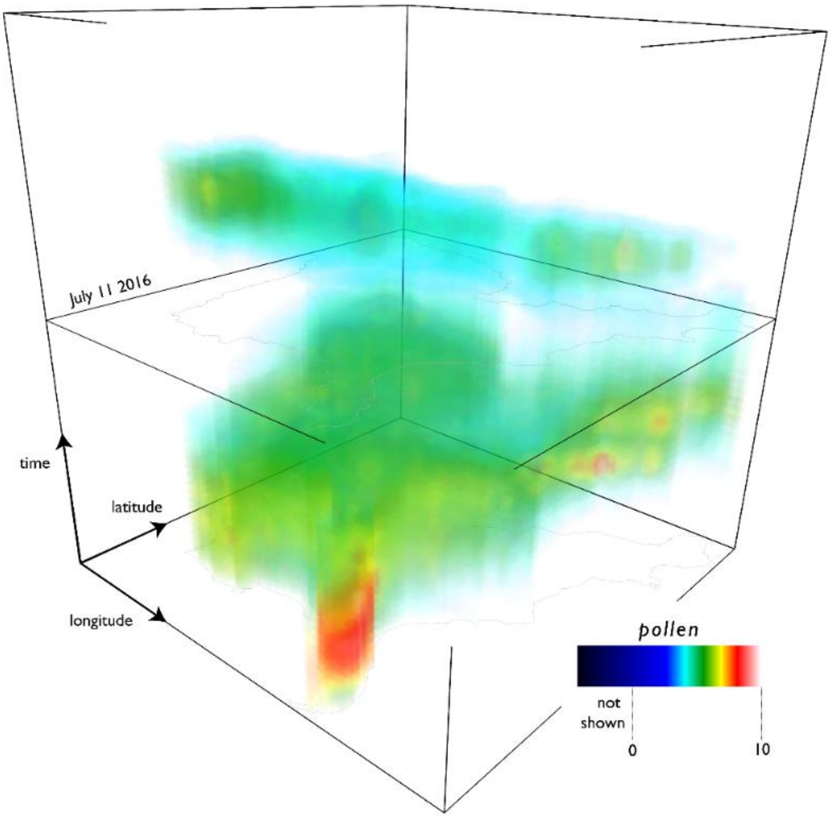


Figure 4: Space-time variation in interpolated pollen values in our study region.

According to our visualization, allergy season substantially diminishes in July until late-summer and early-fall. There is a cluster of high pollen counts in the northeast that occurs in September. There is also a cluster of high pollen counts that occurs in October-November in the southern portion of our study region. This "reverse wedge" is a result of New England plants blooming earlier in the fall because of the potential early snow. Fall blooming plants do not produce the same pollen levels as spring blooming plants, so the number of days between pollen clusters is lower than the clusters seen in the beginning and middle of 2016. Overall, our 3D visualization reconstructs the dynamics of allergy season during 2016 in the Eastern United States.

References

Aguilera, F., et al. 2015. Airborne-pollen maps for olive-growing trees throughout the Mediterranean region: spatio-temporal interpretation. *Aerobiologia* 31: 421-434.

Alba, F., et al. 2006. Airborne-pollen map for *Olea europaea* L. in eastern Andalusia (Spain) using GIS: Estimation models. *Aerobiologia* 22: 109-118.

Cao, X., O. Okhrin, M. Odening, and M. Ritter. 2015. Modelling spatio-temporal variability of temperature *Comput Stat* 30: 745-766.

Charpin, D., and Caillaud, D. 2014. Épidémiologie de l'allergie pollinique. *Revue des Maladies Respiratoires*, 31 (4): 365-374.

DellaValle, C.T., E.W. Triche, and M.L. Bell. 2012. Spatial and temporal modeling of daily pollen concentrations. *Int J Biometeorol* 56: 183-194.

Delmelle, E., Dony, C., Casas, I., Jia, M., & Tang, W. (2014). Visualizing the impact of space-time uncertainties on dengue fever patterns. *International Journal of Geographical Information Science*, 28(5), 1107-1127.

Demšar, U., & Virrantaus, K. (2010). Space–time density of trajectories: exploring spatio-temporal patterns in movement data. *International Journal of Geographical Information Science*, 24(10), 1527-1542.

Ding, Y. and Densham, P.J., 1996. Spatial strategies for parallel spatial modelling. International Journal of Geographical Information Systems, 10(6), pp.669-698.

Fang, T.B., and Y. Lu. 2011. Constructing a Near Real-time Space-time Cube to Depict Urban Ambient Air Pollution Scenario. *Transactions in GIS* 15 (5): 635-649.

Garcia-Mozo, H., C. Galan, and L. Vazquez. 2006. The reliability of geostatistic interpolation in olive field floral phenology. *Aerobiologia* 22: 97-108.

Goovaerts, P. 1997. *Geostatistics for natural resources evaluation*. Oxford University Press on Demand.

Guo, L. et al. 2015. Evaluation of Spatio-Temporal Variogram Models for Mapping Xco_2 Using Satellite Observations: A Case Study in China. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 8 (1): 376-385.

Hohl, A., Delmelle, E., Tang, W., & Casas, I. 2016. Accelerating the discovery of space-time patterns of infectious diseases using parallel computing. *Spatial and Spatio-temporal Epidemiology* 19: 10-20.

Hussain, I., G. Spöck, J. Pilz, and H-L. Yu. 2010. Spatio-temporal interpolation of precipitation during monsoon periods in Pakistan. *Advances in Water Resources* 33: 880-886.

Kyriakidis, P. C., & Journel, A. G. 1999. Geostatistical space–time models: a review. *Mathematical geology* 31 (6): 651-684.

León Ruiz, E. J., García Mozo, H., Domínguez Vilches, E., & Galán, C. 2012. The use of geostatistics in the study of floral phenology of Vulpia geniculata (L.) Link. *The Scientific World Journal.*

Levetin, E., and P.K. Van de Water. 2003. Pollen count forecasting. *Immunology and Allergy Clinics of North America* 23: 423-442.

Li, L., T. Losser, C. Yorke, and R. Piltner. 2014. Fast Inverse Distance Weighting-Based Spatiotemporal Interpolation: A Web-Based Application of Interpolating Daily Fine Particulate Matter PM2.5 in the Contiguous U.S. Using Parallel Programming and k-d Tree. *International Journal of Environmental Research and Public Health* 11: 9101-9141.

Nakaya, T. (2013). Analytical Data Transformations in Space–Time Region: Three Stories of Space Time Cube: Space–Time Integration in Geography and GIScience. *Annals of the Association of American Geographers*, 103: 1100-1106.

Nathan, R. A. 2007. The burden of allergic rhinitis. In *Allergy and Asthma Proceedings* 28 (1): 3-9. OceanSide Publications, Inc.

Pebesma, E.J., K. De Jong, and D. Briggs. 2007. Interactive visualization of uncertain spatial and spatio-temporal data under different scenarios: an air quality example. *International Journal of Geographical Information Science* 21 (5): 515-527.

Rojo, J., and R. Pérez-Badia. 2015. Spatiotemporal analysis of olive flowering using geostatistical techniques. *Science of the Total Environment* 505: 860-869.

Sherman, M. 2011. Spatial statistics and spatio-temporal data: covariance functions and directional properties. John Wiley & Sons.

Siska, P., Bryant, V. M., & Hung, I. 2006. Geospatial analysis of southern pine biome and pollen distribution patterns in Southeastern United States. *GEOGRAFICKY CASOPIS SLOVENSKEJ AKADEMIE VIED* 58 (4): 239.

Siska, P. P., Hung, I., & Bryant Jr, V. M. 2012. The Mapping of Composite Pollen from Point Sampled Data and Cartographic Generalization. *Papers of the Applied Geography Conferences* 35: 191-200.

Wheatley, L.M., and A. Togias. 2015. Allergic Rhinitis. *The New England Journal of Medicine* 372 (5): 456-463.