Voronoi Maps: An Approach to Individual-Based Environmental Exposure Estimation

Wan D. Bae
University of Wisconsin-Stout
baew@uwstout.edu

Shayma Alkobaisi
United Arab Emirates University
shayma.alkobaisi@uaeu.ac.ae

Wade Meyers
Colorado School of Mines
meyerw@scom.math.edu

Sada Narayanappa
Microsoft, USA
sada.narayanappa@microsoft.com

Petr Vojtěchovský
University of Denver
petr@math.du.edu

ABSTRACT

Estimating an individual's environmental exposure is a complicated problem that depends on the amount of time of the individual's exposure, the uncertain location of the individual, and the uncertainty in the levels of environmental factors based on available localized measurements. This problem is critical in the applications of environmental science and public health. In this paper we study the fundamental issues related to spatio-temporal uncertainty of human trajectories and environmental measurements and define a model of exposure uncertainty. We adopt a geometric data structure called the Voronoi diagram to interpolate environmental data, and utilize it in our proposed method to efficiently solve this problem. We evaluate the performance of the proposed method through experiments on both synthetic and real road networks. The experimental results show that our solution based on probabilistic routing aggregation is an efficient and extensible method for environmental exposure time estimation.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Spatial databases and GIS

Keywords

exposome, environmental exposure, trajectory uncertainty, mobile sensors, individual-based healthcare

1. INTRODUCTION

Relations between negative health effects like lung cancer and asthma and elevated levels of the environmental factors, such as air pollution and tobacco smoke, have been detected in several large scale exposure studies [12, 24]. It has been also shown that the combination of extreme weather conditions, such as heat and high humidity, escalates some diseases' episodes and consequently leads to mortality [1]. In many of these studies environmental triggers related to human health have been measured in a broad manner, (i.e., studies were based on summarized data collected in large scale areas such as cities). However, individual correlations may differ significantly from those of the population level due to individual behavior and spatial and temporal variability of the environment [13]. Thus, individual-level measurement of environmental exposure is important for developing more accurate diagnoses of the causes of diseases.

The study of environmental exposures requires synchronization of environmental data, individuals' moving trajectories, and the behaviors of individuals, such as enroute activities, route selections, etc. By incorporating weather and other environmental data to characterize and predict the route distribution (path selection behavior), one can measure an individual's exposure to certain environmental conditions. Moreover, with the use of spatial mining tools to evaluate individual's exposure, one can identify environmental triggers, predict behavior of individual exposure and potential risk, and provide optimal intervention. A model overview of this complete individual-based health intervention is shown in Figure 1.

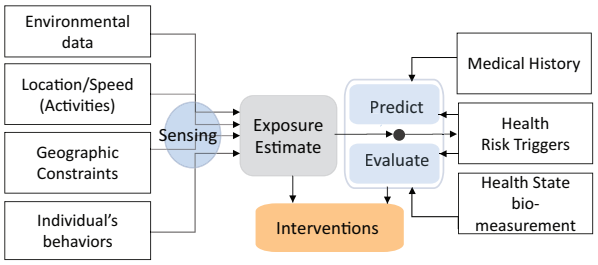


Figure 1: Model overview

Health intervention systems are expected to provide individuals with real-time responses based on predictive exposure estimates. However, there exist several challenges in aggregating spatio-temporal data which are voluminous in nature. First, one cannot realistically sample too frequently in order to keep data amounts reasonable and sensors long lasting [9, 14]. Second, a challenge exists in the spatio-temporal uncertainty associated with the sampled data. In addition, the join operation among spatio-temporal datasets in the presence of multi-level uncertainties is extraordinarily time-consuming and data-intensive. Finally, behavior of

SAC 2016, April 04-08, 2016, Pisa, Italy

DOI: http://dx.doi.org/10.1145/2851613.2851715

individuals has a large impact on exposure estimates, and such behavior is not known *a priori*, but rather built iteratively from previous estimates of exposure and individuals' responses.

In this paper, we introduce a novel model to define environmental exposure uncertainty and propose a method that copes with uncertainty. Our proposed method constructs simple and memory-efficient environmental maps and joins the maps with a set of possible routes (paths). Exposure times of these routes are aggregated using probabilities assigned to the routes.

2. RELATED WORK

Recent advances in mobile sensor, communication and computation have opened new opportunities to investigate the relationship between pollution, human behavior and health outcomes as well as optimized interventions [22]. The authors in [22] developed an air pollution monitoring system that provides real-time interpolated maps of air quality using GIS, which analyzes and displays data collected by sensors. Such successful deployment of mobile devices for environmental monitoring have led researchers to use them in health assistant systems like the "EnviroFlash" model [2] that is used to notify patients about up to date information on air quality. However, despite isolated commercial successes [2, 20], a realization of these opportunities requires us to surmount considerable computing challenges involving estimation of exposure time combined with human spatial behavior.

To simplify the relationship between stochastic, spatio-temporal sequences of pollutant concentration and their physiological consequences, researchers have recently began to entertain the notion of "exposome" [24]. Most recently, exposome has been defined as "The cumulative measure of environmental exposures (influences) and associated biological responses throughout the lifespan, including exposures from the environment, diet, behavior, and endogenous processes" [21]. Different tools and approaches exist to monitor environmental exposures such as those proposed in [24].

Spatial and spatio-temporal data analysis methods [7] together with individuals' electronic health records provide ways of relating health and human disease to specific genetic and environmental factors [8]. The density, accuracy, and specificity of current geospatial data also facilitate sophisticated spatial and spatio-temporal analysis and the modeling of complex spatial health processes at the level of the individual rather than the aggregate [18, 19]. Research has shown that even at the population level of environmental exposure, the results can vary based on the technique used [16]. In that paper, the authors discussed the tradeoff between computational cost and accuracy of exposure measurement.

Some research work in the area of geoscience presented modeling methods using GPS trajectory data for an individual's exposure estimation [10, 11]. A model presented in [10] was used to disaggregate air quality to quantify the individual's exposure to air pollution. A case study for 10 individuals was conducted in Germany. The authors assumed that GPS time/location data was collected every 1 second and focused on disaggregate concentration values at the locations in a cell.

None of the aforementioned approaches takes location uncertainty as a factor when modeling and measuring individual's exposure. They all assume known locations of individuals at a scale of 1-5 seconds. However, due to computational limitations as well as sensor and mobile energy efficiency, our proposed approach assumes longer periods of gap in reporting individual's locations, and, hence, taking uncertainty in the individual's location into consideration when measuring exposure. In addition, our model provides a more refined representation of environmental concentrations through proposing a Voronoi diagram-based structure called the Voronoi map, which is more efficient compared to heat maps used in previously proposed solutions.

3. EXPOSURE UNCERTAINTY MODELING

3.1 Uncertainties of Spatio-temporal Data

Individuals' trajectory data is often used in the assessment of risk factors due to environmental exposures by integrating them with GIS data involving the spatio-temporal variations of these risk factors [19]. Trajectories representing the location and time of individuals are typically obtained through sampling using GPS and they are stored as a set of points or lines (vector data format). Although GPS can capture an individual's movement or location changes every second, it would require very large amount of space and computational resources to process. Thus, an individual's continuously changing properties are typically discretely updated (sampled), for instance in intervals ranging from tens of seconds to tens of minutes. Hence, individuals' trajectories are always associated with a degree of uncertainty, especially when there is a considerable time gap between two updated values [23].

Environmental data is interpolated and stored in a heat map, which is normally presented in raster format. The concentration of an environmental factor is a spatio-temporal variable represented by grid cells for discrete timestamps, with the grid cell value representing a spatially and temporally averaged estimate of the true concentration. The scale of air quality models is usually a few hundreds of meters and the data is normally captured every 30-60 minutes. Measuring such variables is associated with spatial and temporal uncertainties because of the approximations and interpolations used in modeling [10, 11, 15, 17].

In order to calculate an individual's exposure to environmental factors, a join operation is executed among the trajectory dataset of the individual and the environmental datasets. However, little research has been done on the optimization or approximation techniques for spatial joins on datasets presented in different data formats. Figure 2 illustrates trajectories in vector format of two individuals overlaid on the environmental dataset represented in raster format.

A simple way to perform this join is to convert the trajectory dataset into raster format and then overlay the trajectories on an environmental dataset to calculate their overlaps (the location and the length/area where they overlap). However, the overlay operation to estimate individuals' exposure time can be complicated (because of the presence of multiple uncertainties) and time consuming (because of multi-joins on spatio-temporal datasets).

3.2 Exposure Uncertainty Model

This paper presents a novel exposure uncertainty model for individuals whose movements are based on road networks.

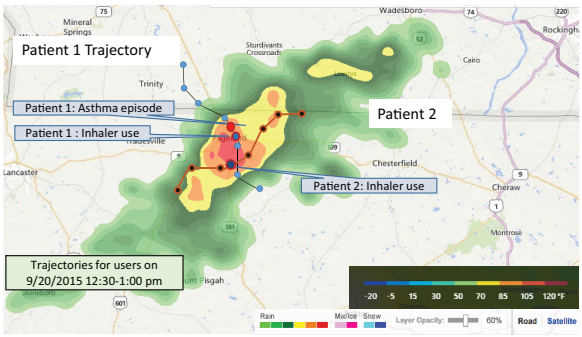


Figure 2: Uncertain trajectories on a heat map interpolated from weather stations

The formula for the exposure, E_i, to environmental factor i given by this model is: $E_i = \sum_{j=1}^{n} T_{ij} * C_{ij}$, where T_{ij} is the time spent in a region j by the individual and C_{ij} is the air pollutant concentration of the environmental factor i that the individual is exposed to in region j. In order to deal with the spatio-temporal uncertainty of an individual for time interval t between two consecutively reported GPS locations, we will first need to estimate T_{ij}, the *exposure time*, for each i and j.

Our proposed model is based on a probability distribution function that represents the location uncertainty of a moving individual in a road network. The problem can be defined as follows: Let $G = (V, E)$ be a graph in $\mathbb{R}^2$ representing a road network, where V and E are the set of vertices and edges. Each edge represents a road segment and has attributes, length l and maximum speed s allowed on the edge. Then a trajectory $I_{t_1 t_n}$ of a moving object in road network G is a sequence of points (positions) with time stamps: $I_{t_1 t_n} = \{(t_1, p_1), (t_2, p_2), .., (t_n, p_n)\}$, where p_m is the sample point (position) at t_m, for $m = 1, 2, 3, .., n$, and $t_{m+1} > t_m$.

Given two sampled points p_m and p_{m+1} of an object's trajectory, let $R_m = \{r_1, r_2, .., r_k\}$ be the set of the candidate routes (i.e., k top rank routes which the individual is most likely to choose) between p_m and p_{m+1}. The candidate routes are selected from all possible routes between p_m and p_{m+1}, where the object possibly walked (drove) in a road network G with a given time interval $t = t_{m+1} - t_m$, and a set of moving restrictions (velocity v and/or acceleration a). If the probability density function of selecting a route r_j was uniform, then the probability of one of the possible routes would be: $P_{r_j} = \frac{1}{|R_m|} = \frac{1}{k}$, where k is the cardinality of the set R_m (the number of candidate routes).

A different probability is assigned to each route for exposure time calculation. We then have a route r_j as a set of the edges, $r_j = \{r_{j1}, r_{j2}, .., r_{js}\}$. Given that the exposure time to a particular environmental data value is proportional to the ratio $|r_{jh}|/|r_j|$, where r_{jh} is the path segment, then we have the following equation:

$$T_{r_{jh}} = \frac{|r_{jh}|}{|r_j|} * t * P_{rj}. \quad (1)$$

In future work, this rather simple assumption must be replaced by a more realistic approach based on the behavior of the object, yielding more accurate probability density functions (PDFs). The PDFs would incorporate behavioral aspects of individuals and/or statistical information about road usage.

Applying exposure time described in Equation (1), an estimate of exposure to environmental factor i with the given time interval t is given by: $\bar{E}_{iR_m} = \sum_{j=1}^{k} \sum_{h=1}^{s} T_{r_{jh}} * C_{r_{jh}}$. Finally, the exposure estimation along the trajectory $I_{t_1 t_n}$ can be calculated as follows:

$$\bar{E}_i = \sum_{m=1}^{n} \bar{E}_{iR_m}. \quad (2)$$

4. THE VORONOI MAP: ENVIRONMENTAL DATA INTERPOLATION

Currently there are 41446 weather stations in the databases from across various administrative divisions of the U.S, including states and territories. These weather stations update some environmental data every 30 minutes (or 60 minutes) and these data files are publicly accessible through the National Centers for Environmental Information and the National Oceanic and Atmospheric Administration [3].

Recent research work in environmental science and geoscience have provided methods for interpolation of discrete environmental data to generate environmental heat maps. Generating heat maps of environmental data and storing the maps every 30-60 minutes is a data-intensive and time consuming process. Moreover, the "raster overlay" operation between trajectories and multiple environmental heat maps is very challenging in terms of space and time complexity, particularly in the presence of uncertainties. Hence, a more robust method for capturing environmental measurements and storing the data in an efficient way is needed.

The Voronoi diagram is the partitioning of a plane with n points into convex polygons such that each polygon contains exactly one generating point and every point in a given polygon is closer to its generating point than to any other generating point [6]. Voronoi diagram and its dual graph, Delaunay triangulation, are used in many applications in areas such as computer graphics, epidemiology, geophysics, and meteorology. Delaunay triangulations maximize the minimum angle of all the angles of the triangles in the triangulation and hence they tend to avoid skinny triangles.

In this paper, we propose a model called the "Voronoi map" to store environmental data. Instead of interpolating the raw data values of environmental datasets into multiple heat maps whenever the data is updated, we create a "Voronoi map" using the locations of weather stations for real-time exposure estimation. Closer environmental stations would provide better estimations, so we use Voronoi cells to determine which station should be chosen to provide the environmental measurement for a specific individual based on his/her location. Each Voronoi cell in the map represents a region (area) that has similar environmental conditions. The values of environmental datasets are stored as properties of the cells in the Voronoi map.

When weather stations are close to each other, Voronoi cells tend to be small and the Voronoi cell values can accurately describe the environmental condition in the cell. On the other hand, the cell values can start being inaccurate when weather stations are further apart, such as in rural areas. Also, the cell values can be less accurate moving away from the station and close to the edge of the cell. To solve this problem, we propose a method that refines the Voronoi map through multi-level Delaunay triangulations.

Our proposed method uses interpolation of the cell values from the weather stations. An example of the refinement steps is shown in Figure 3. Let VD_i and DT_i be a set

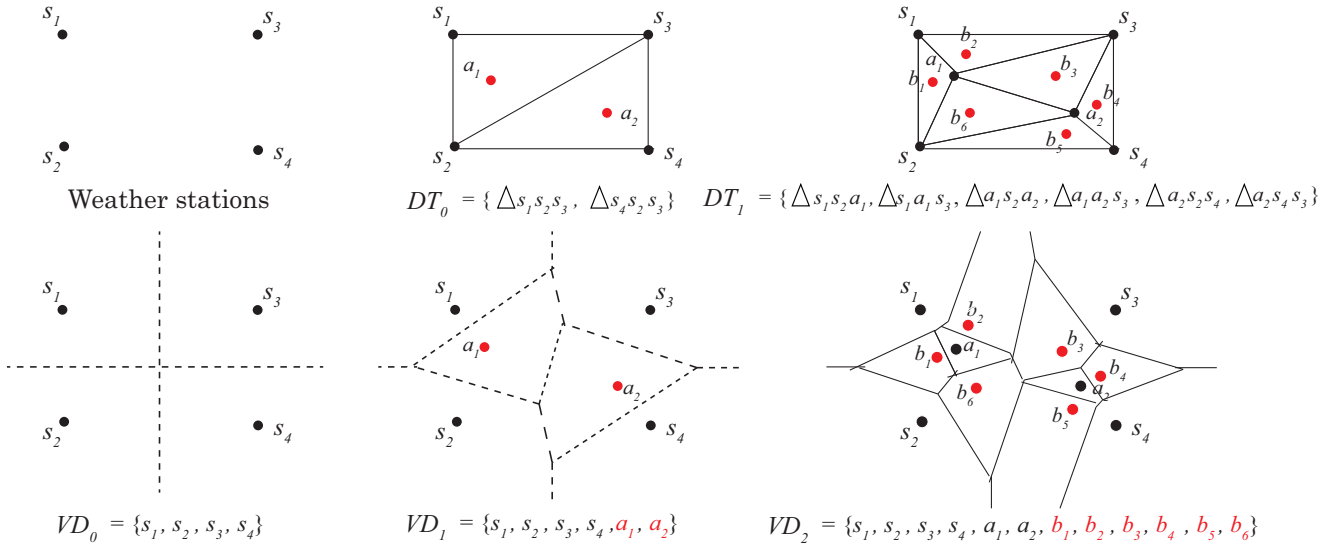


Figure 3: Construction of the Voronoi Map

of points in a Voronoi diagram and a set of triangles in a Delaunay triangulation at level i, respectively. Let point s_i represent the weather station i with all properties of the weather station i stored in s_i. The Voronoi diagram at level 0, VD_0, includes the weather station points s_1 - s_4, as shown in Figure 3. The value of each cell in VD_0 where point s_i lies is an environmental data value from weather station s_i. We then create DT_0, the Delaunay triangulation at level 0 using VD_0. For the level 1 refinement step, we find the center points of all triangles in DT_0 and add these points (a_1 and a_2) into the current Voronoi diagram VD_0, we then create a new Voronoi diagram VD_1. The environmental data values stored in these center points are calculated using the values stored in the three points of the corresponding triangle. For example, the value of a_1 can be calculated from the interpolation of the values stored in the three points of $\triangle_{s_1 s_2 s_3}$; Let l_1, l_2, and l_3 be the lengths of $\overline{a_1 s_1}$, $\overline{a_1 s_2}$, and $\overline{a_1 s_3}$, respectively. Let l be the sum of l_1, l_2, and l_3, and e_1, e_2, and let e_3 be the environmental values stored in s_1, s_2, and s_3, respectively. We then calculate $t_1 = \frac{l-l_1}{l} * e_1$, $t_2 = \frac{l-l_2}{l} * e_2$, and $t_3 = \frac{l-l_3}{l} * e_3$, and finally we calculate the value $a_1 = \frac{t_1+t_2+t_3}{3}$. For the level 2 refinement step, we find the center points of all triangles in DT_1 and add them to VD_1 to obtain a new Voronoi diagram VD_2 and this is followed by calculation of the values of newly found center points. The refinement step continues until i reaches a desired granularity level k while updating VD and DT. Finally, we create a Voronoi map using VD_k (k is set to 2 in our example). This incremental construction supports more accurate environmental estimates since as we keep refining we can further provide more accurate interpolation.

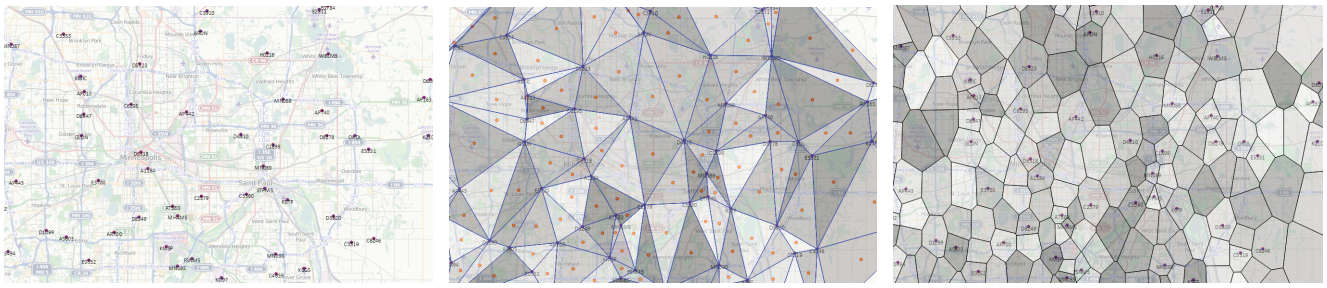
The next step in our exposure measurement method is to join the captured environmental data with the individual's trajectory to calculate the exposure of the individual. Since the weather stations are static (fixed locations), the Voronoi map is created once for a trajectory dataset and there is no need to update the structure. Instead the properties of each cell in the Voronoi map are updated every 30-60 minutes with new data values received from the weather stations. In order to enhance the process of calculating the exposure time, we can precompute the points of intersection between a road network and Voronoi diagrams. An example in Figure 4 demonstrates the use of the Voronoi map in our exposure estimation method.

Although the idea is simple, the proposed method is powerful with the following advantages: (1) the creation of the Voronoi map for environmental datasets is more computationally efficient compared to existing methods that use traditional interpolation, (2) the cells in a Voronoi map store the values from multiple environmental datasets, hence storing data using this geometric data structure is more efficient in space than storing each raster cell values in multiple heat maps, (3) the join operation between trajectory data (lines) and the cells (polygons) of the Voronoi map is less time consuming than raster overlay between trajectory data and heat maps.

5. PROBABILISTIC EXPOSURE TIME AGGREGATION

As a part of the environmental exposure model for individuals developed here, we propose and test a probabilistic method for exposure time aggregation. This method joins the individual's trajectory dataset (a set of discrete location/time values) and Vornoi maps constructed using environment datasets from weather stations. It calculates an estimated exposure time for each of the possible paths and aggregates these values to compute a final estimation.

First, we calculate the length of a path within a Voronoi cell (region). This length can reveal exposure time by multiplying it by travel speed. We assume that we know ahead of time the actual exposure length of the path so that we can compare our estimates to the actual result. We limit the possible paths taken to a small fixed number. In practice this can be done using the time interval, and speed data (velocity) to get practical path options. In our experiments, we use the shortest paths calculated by Dijkstra's k-shortest path algorithm implemented in *pgRouting* [5]. We assume a fixed velocity during a path for simplicity.



(a) Weather stations in Minnesota (b) Delaunay triangulation (c) Voronoi map

Figure 4: Voronoi weather map

5.1 Road Networks on the Voronoi Map

To show the steps of the method, we use both a real road network with real weather stations and a synthetic road network with synthetic weather stations. For the first example of the real road network, the Voronoi map cells (weather regions) are created based on the real weather stations' locations illustrated in Figure 5. The weather conditions, i.e., temperature, humidity, ozone, etc., in the green colored region are from weather station $C6295$, the weather conditions in the yellow colored region are from weather station $AP442$, and the weather conditions in the mauve colored region are from weather station $D6318$.

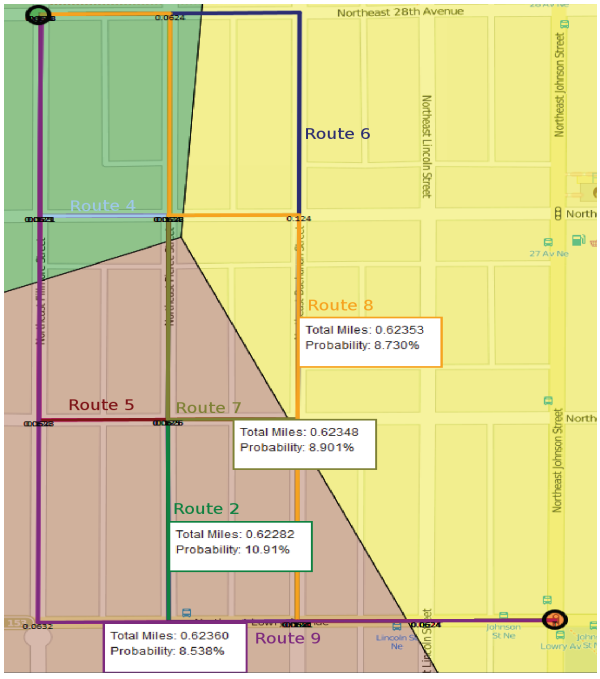


Figure 5: The real road network with real weather stations

In the examples we use two given points (two reported locations) to analyze the possible routes between the two points, the green filled circle at the beginning of the route in the top left and the orange filled circle ending the route in the bottom right in Figure 5, and then the exposure time along the routes is calculated. Based on the possible routes between the individual's two reported points, what is the individual's exposure time to each of the weather regions? To answer this, we may solve a similar problem which is to calculate the distance traveled in each region.

The synthetic road networks and synthetic weather stations in Figure 6 (a) are set up to model the real version of the exposure estimation. The subregion has 5 roads going north and south and 5 roads going east and west. Assuming that 4 weather stations are located in this region, a Voronoi map is generated and synthetic environmental values are stored as the Voronoi cell properties.

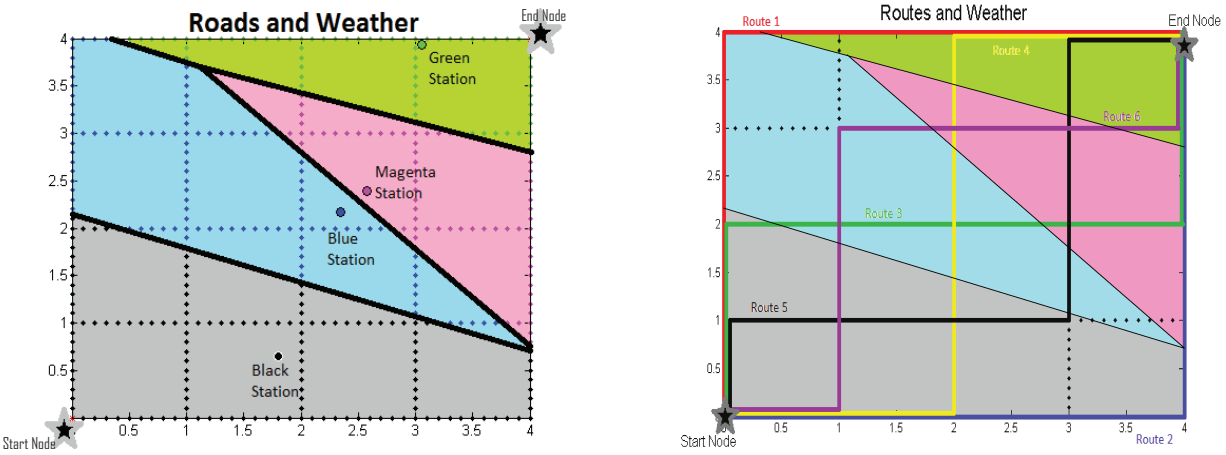
5.2 Possible Routes

Since we would not know the actual route ahead of time, we first need to calculate some possible routes between the beginning and end points. This could be done in practice by using the elapsed time between start and end points, and realistic speed along paths.

For the real road network we consider the 10 shortest routes using *pgRouting* as possible routes as shown in Table 1. The reason for choosing 10 routes as candidate routes selected by an individual is merely to validate the approach and does not necessarily indicate sufficiency.

route	total length	length in AP442	length in C6295	length in D6318
Route 0 (blue)	0.6224	0.07148	0.2019	0.3490
Route 1 (fuchsia)	0.6226	0.07148	0.1642	0.3870
Route 2 (green)	0.6228	0.07148	0.2023	0.3490
Route 3 (#800000)	0.6230	0.08611	0.2019	0.3350
Route 4 (#99CCFF)	0.6231	0.2627	0.1930	0.1674
Route 5 (maroon)	0.6232	0.08611	0.1642	0.3730
Route 6 (navy)	0.6234	0.3785	0.07749	0.1674
Route 7 (olive)	0.6235	0.08611	0.2023	0.3350
Route 8 (orange)	0.6235	0.2627	0.1934	0.1674
Route 9 (purple)	0.6236	0.07148	0.1642	0.3880

Table 1: Possible routes in the Real Road Network



(a) The synthetic road networks with synthetic weather (b) The synthetic routes used

Figure 6: Synthetic road network with synthetic weather

For the synthetic road network we consider the routes shown in Figure 6 (b). Each of the routes represents a possible path that could have been taken to go from the start point in the bottom left to the end point in the top right.

5.3 Probabilities on Routes

At this point, we would like to calculate and assign a probability to each route based on how likely it is to be the actual route. In other words, what are the chances that the individual took a specific route to get from the point of origin to the target point. As a proof of concept, we consider two cases for the probabilities; in one case we use equal probabilities; in the example of the real road network, since we have 10 routes, each route has a 10% chance of being the actual route taken by the individual. In the other case, we use non-uniform probabilities.

One way to more accurately predict the exposure time to the weather regions is to know which paths are more likely to be taken. Assigning probabilities for each route can be done in different ways, but for simplicity, in this experiment we use the path length to determine its likelihood of being chosen by an individual. We want the longer paths to have lower probabilities of being taken, while shorter paths would be more likely to be taken. The method of assignment of probabilities exploits the fact that humans generally prefer shorter routes when traveling from one location to another. Thus, when we calculate the exposure estimation, the shorter routes are weighted more heavily, thus increasing the accuracy of our estimate if one of the shorter routes turn out to be the actual route taken by the individual.

It is worth noting that a more accurate distribution function that models how humans select routes and how their behaviors affect their selection, would enhance our method and would make the results of our experiments more realistic. For the exposure estimates, it would be also important to consider levels of en-route activity, such as riding a vehicle, strolling, walking or running, which would affect the rate of breathing and thereby intake of the pollutant. A more "individualized" model that would take human behavior as an input as well as other possible factors such as traffic will be considered in future work.

The following calculation is used to come up with each path probability based on its length. Let P_i be the probability of route r_i being taken, L_i be the length of r_i, D_i be the difference between the length of that path and the average length of all routes and k be the desired number of possible routes. By taking the possible k routes from P_0 to P_k, we calculate $\bar{x}$ which is the average length of all the routes:

$$\bar{x} = \frac{L_0 + L_1 + ... + L_k}{N}$$

Now we calculate $D_0 - D_k$:

$$D_i = \bar{x} - L_i$$

Once we have $D_0 - D_k$ we know which routes are longer than the average and shorter than the average and by how much. Routes that are longer than the average will have a negative D_i, and routes that are shorter than the average will have a positive D_i. We then calculate the absolute total of these values. Let this be

$$s = |D_0| + |D_1| + ... + |D_k|,$$

Now we can get the route probabilities.

$$P_i = \frac{1}{N}\left(1 + \frac{D_i}{s}\right) \quad (3)$$

The value k in our example in the real road network is 10. Therefore, $P_0 = 12.22\%$, $P_1 = 11.61\%$, $P_2 = 10.91\%$, $P_3 = 10.21\%$, $P_4 = 10.04\%$, $P_5 = 9.605\%$, $P_6 = 9.227\%$, $P_7 = 8.901\%$, $P_8 = 8.730\%$, and $P_9 = 8.538\%$

Using the probability for route r_i, we calculate an estimate of the exposure time for r_i, $i = 0, 1, 2, ..., k-1$. All estimated exposures are aggregated to a final estimation of the individual's exposure time between the start point and the end point.

6. PERFORMANCE EVALUATION

To compare how probabilities of routes affect the exposure time estimation, we choose one of the routes to represent the actual route and repeat the experiments so that each time a different route is selected as the actual route.

6.1 Even Probabilities

We apply even probability on each route of the possible routes for the real road network with real weather stations. In our experiment, Route 1 is selected as the actual route with the following properties:

Route 1 (color: fuchsia)	
Probability	10.00%
Total length	0.6226
Length in station AP442	0.0715
Length in station C6295	0.1642
Length in station D6318	0.3870

For each route, we use the length of the route overlapped with each cell of the Voronoi map along with the path probability to calculate the estimated exposure time. Recall that the probability of each route in this case is $\frac{1}{10}$. Let $\hat{w}_1$ be the estimated exposure length to weather station AP442, $\hat{w}_2$ be the estimated exposure length to weather station C6295, and $\hat{w}_3$ be the estimated exposure length to weather station D6318. Then we have, $\hat{w}_1 = 0.1448$ miles, $\hat{w}_2 = 0.1765$ miles, and $\hat{w}_3 = 0.3018$ miles. Table 2 shows the result of the experiment. With even probabilities our estimate was off by a total of 0.17082.

Even probability exposure lengths			
Weather station	Estimated	Actual	Absolute error
AP442	0.1448	0.0715	0.0733
C6295	0.1765	0.1642	0.0123
D6318	0.3018	0.3870	0.0852

Table 2: Experiment on real road network with even probabilities and actual route being Route 1

6.2 Varied Probabilities

With varied probabilities shown in Table 3, we have the following values for the estimated exposure lengths to weather stations AP442, C6295, and D6318, respectively: $\hat{w}_1 = 0.1399$ miles, $\hat{w}_2 = 0.1776$ miles, and $\hat{w}_3 = 0.3055$ miles. In the experiment on the real road network with varied probabilities where Route 1 is selected as the actual route, the estimate was off by a total of 0.16332. This is slightly better than the result of the even probabilities case. This is because the actual route is one of the more probable routes. As expected, more accurate probabilities assigned leads to more accurate prediction of exposure.

Exposure lengths with varied probability			
Weather station	Estimated	Actual	Absolute error
AP442	0.1399	0.0715	0.0684
C6295	0.1776	0.1642	0.0134
D6318	0.3055	0.3870	0.0815

Table 3: Experiment on real road network with varied probabilities and actual route being 1

One way to confirm this is to choose each route to be the actual route and see which cases have the most accurate exposure estimation. The results of the real road network case are shown in Table 4. In the higher probability routes the error tends to remain lower. This makes sense since more probable routes are weighted higher when calculating the estimation.

Varied actual routes		
Route	Probability	Absolute error
0	0.1222	0.13622
1	0.1161	0.16332
2	0.1091	0.13662
3	0.1021	0.10759
4	0.1004	0.2763
5	0.09605	0.13469
6	0.09227	0.47681
7	0.08901	0.10799
8	0.08730	0.2767
9	0.08538	0.16432

Table 4: Experiment on real road network with varied probabilities and considering all routes being the actual route

In the experiment for the synthetic road network with synthetic weather stations, each route is selected as the actual route and the accuracy of the estimates depending on the probability of the route selection is evaluated. We calculate the error for each route; meaning we allow each route to be considered the actual route and then see what the error would have been if that was the actual route. For example, if Route 6 was the actual route, we compare our estimate to Route 6 to find how close our estimate was. We ran this many times while randomly picking locations for the weather stations.

The error calculation is done by taking the estimated exposure time and subtracting the exposure time for that route, then taking the absolute value of the result. The test result of the synthetic road and weather are shown in Figure 7. Note that the error could be a maximum of 162 seconds. This does not seem obvious since the route is only 81 seconds long, but, for example, if the estimate was 81 seconds in the green region but the actual route was 81 seconds in the black region that would evaluate to 162 seconds of error. Table 5 shows the results from running the experiment 500 times.

Synthetic route/weather experiments		
Route	average error (sec)	standard deviation (sec)
1	56.0367	14.2981
2	55.8893	13.6654
3	32.5733	13.4831
4	33.276	13.6131
5	23.8093	10.5005
6	24.8493	10.8699

Table 5: Data from 500 iterations of the synthetic route and weather experiment

6.3 System Implementation

The following technologies are used to implement our evaluation system: The system uses Linux Redhat 6.0 operating system; Java and Python are the main programming languages; Javascript is used for most of the client systems. The system implements data repository using Hadoop file system and PostGres database. Third party data visualization tool such as OpenLayers [4] is used to create base maps and integrate geographical datasets. The client codes that invoke Representational State Transfer (REST) API to pull results

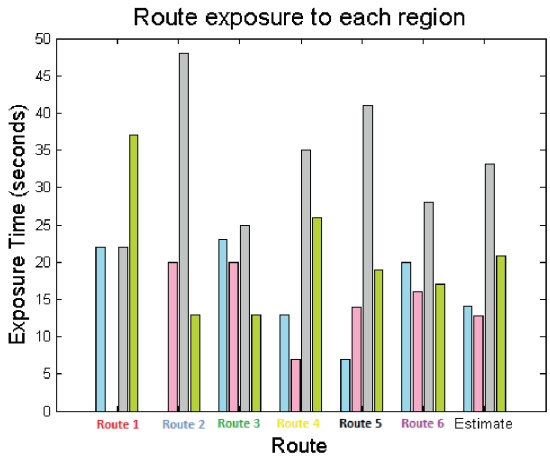


Figure 7: Each routes actual exposure time to each weather region along with the estimated exposure times

from its data repository are written in Javascript/AJAX, and are integrated with OpenLayers for the user interface. The Apache Tomcat is used as an application server.

7. CONCLUSIONS

Exposure measurement is the key to explain the possible effects of environmental conditions on humans. In this paper, we proposed a method to calculate exposure time estimation at an individual level and in the presence of location uncertainty. Our method takes advantage of the simple and fast construction of Voronoi diagrams to build a new structure called Voronoi Maps, which models the environmental factors in refined cells at different granularity levels. We then apply probabilistic functions to determine the exposure time of individuals to environmental factors. Experiments on synthetic and real datasets showed that our proposed solution is robust and efficient.

8. ACKNOWLEDGMENTS

This material is based upon works supported in part by the Information and Communication Technology Fund of United Arab Emirates under award number 21T042 and in part by the National Science Foundation under award number CNIC-1338378.

9. REFERENCES

[1] Centers for disease control and prevention. `http://www.cdc.gov/asthma`.

[2] Environflash. `http://www.enviroflash.info`. Accessed: August 2015.

[3] National centers for environmental information and the national oceanic and atmospheric administration. `http://www.ncdc.noaa.gov/cdo-web`.

[4] Openlayers. `http://openlayers.org/`.

[5] pgrouting. `http://pgrouting.org`.

[6] F. Aurenhammer. Voronoi diagramsâĂŤa survey of a fundamental geometric data structure. *ACM Computing Surveys (CSUR)*, 23(3):345–405, 1991.

[7] N. Cressie and C. K. Wikle. *Statistics for spatio-temporal data*. John Wiley & Sons, 2011.

[8] R. Feil and M. F. Fraga. Epigenetics and the environment: emerging patterns and implications. *Nature Reviews Genetics*, 13(2):97–109, 2012.

[9] B. Gedik and L. Liu. Mobieyes: Distributed processing of continuously moving queries on moving objects in a mobile system. In *Advances in Database Technology - EDBT 2004*, pages 67–87. 2004.

[10] L. Gerharz and E. Pebesma. Using geostatistical simulation to disaggregate air quality model results for individual exposure estimation on gps tracks. *Stochastic Environmental Research and Risk Assessment*, 27(1):223–234, 2013.

[11] L. E. Gerharz, A. KrÃijger, and O. Klemm. Applying indoor and outdoor modeling techniques to estimate individual exposure to pm2.5 from personal {GPS} profiles and diaries: A pilot study. *Science of The Total Environment*, 407(18):5184–5193, 2009.

[12] J. M. Gibson, A. Brammer, C. Davidson, T. Folley, F. Launay, and J. Thomsen. *Environmental Burden of Disease Assessment*. Springer, 2013.

[13] C. A. Gotway and L. J. Young. Combining incompatible spatila data. *Journal of the American Statistical Association*, 97(458):632–648, 2002.

[14] R. H. Güting and M. Schneider. *Moving objects databases*. Elsevier, 2005.

[15] T. Hägerstraand. What about people in regional science? *Papers in regional science*, 24(1):7–24, 1970.

[16] I. Hanigan, G. Hall, and K. B. Dear. A comparison of methods for calculating population exposure estimates of daily weather for health research. *International Journal of Health Geographics*, 5(1):38, 2006.

[17] G. B. Heuvelink, J. D. Brown, and E. Van Loon. A probabilistic framework for representing and simulating uncertain environmental variables. *International Journal of Geographical Information Science*, 21(5):497–513, 2007.

[18] M.-P. Kwan. From place-based to people-based exposure measures. *Social science & medicine*, 69(9):1311–1313, 2009.

[19] M.-P. Kwan. How gis can help address the uncertain geographic context problem in social science research. *Annals of GIS*, 18(4):245–255, 2012.

[20] N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. T. Campbell. A survey of mobile phone sensing. *Communications Magazine, IEEE*, 48(9):140–150, 2010.

[21] G. W. Miller and D. P. Jones. The nature of nature: refining the definition of the exposome. *Toxicological Sciences*, 137(1):1–2, 2014.

[22] O. Pummakarnchana, N. Tripathi, and J. Dutta. Air pollution monitoring and gis modeling: a new use of nanotechnology based solid state gas sensors. *Science and Technology of Advanced Materials*, 6(3):251–255, 2005.

[23] S. Schappert and E. Rechtsteiner. Ambulatory medical care utilization estimates for 2007. *Vital and Health Statistics. Series 13, Data from the National Health Survey*, (169):1–38, 2011.

[24] C. P. Wild. The exposome: from concept to utility. *International journal of epidemiology*, 41(1):24–32, 2012.