

# Case-control clustering for residential histories

Geoffrey M. Jacquez<sup>1</sup>, Andy Kaufmann<sup>1</sup>, Jaymie Meliker<sup>2</sup>, Pierre Goovaerts<sup>1</sup>, Gillian AvRuskin<sup>1</sup> and Jerome Nriagu<sup>2</sup>

1 BioMedware, Inc., 516 North State Street, Ann Arbor, MI

2 The University of Michigan School of Public Health

## **Abstract**

**Background:** This paper introduces a new approach for evaluating clustering in case-control data that accounts for residential histories. Although many statistics have been proposed for assessing local, focused and global clustering in health outcomes, few, if any, exist for evaluating clusters when individuals are mobile.

**Methods:** Local, global and focused tests for residential histories are developed based on sets of matrices of nearest neighbor relationships that reflect the changing topology of cases and controls. Exposure traces are defined that account for the latency between exposure and disease manifestation, and that use exposure windows of varying duration. Several of the methods so derived are applied to evaluate clustering of residential histories in a case-control study of bladder cancer in south eastern Michigan.

**Results** Statistically significant clustering of residential histories of cases was found but is likely due to delayed reporting of cases by the University of Michigan hospitals.

**Conclusions** Because humans are mobile, the methods proposed in this paper are preferred over traditional approaches that assume sessile individuals.

## **Background**

U.S. population-based surveys estimate that adults spend 87% of their day indoors, 69% in their place of residence, and 6% in a vehicle (Collia et al 2001, Klepeis et al 2001, Reuscher et al 2002). To date, almost all published disease cluster investigations use static geographies in which individuals are assumed to be sessile (e.g. geocoded place of residence at time of diagnosis or death is used to record the locations of health events), even though most researchers acknowledge that residential mobility should be accounted for, especially for diseases with long latencies such as cancer. In a recent review of standard methods for evaluating exposure/hazards, disease mapping and clustering techniques, Bayesian approaches, Markov Chain Monte Carlo (MCMC) and geostatistical methods, Mather et al (2004) identify as substantial weaknesses (1) the lack of temporal referencing of geospatial data and (2) the inability of methods to account for residential histories. A recent meeting of this nation's experts recognized the need to account for latency and human mobility as especially pressing in studies of cancer (Pickle, Waller and Lawson 2004).

The representation of individuals as sessile (immobile) rather than vagile (mobile) in part is due to the static world view of GIS software, which is largely incapable of representing both human mobility and temporal change (Goodchild 2000). Recently, technological advances have resulted in Space Time Intelligence Systems (e.g. Jacquez et al 2005; Greiling et al 2005; Meliker et al 2005) that implement several constructs from Geographic Information Science for representing human mobility (see Miller 2004 for a review). The methods presented in this paper build on this body of prior work to produce case-control cluster statistics for residential histories.

We begin with a brief background on tests for disease clustering, followed by a summary of approaches to modeling human mobility. We then develop a suite of novel tests for evaluating local, global and focused clustering in residential histories using case-control data. Finally, we illustrate the new techniques by quantifying local and global clustering of residential histories in a case-control study of bladder cancer in Michigan.

## **Background on cluster tests**

Cluster tests work within a hypothesis testing framework that proceeds by calculating a statistic (e.g. clustering metric) to quantify a relevant aspect of spatial pattern in a health outcome (e.g. case/control location, disease incidence, or mortality rate). The numerical value of this statistic is then compared to the distribution of that statistic's value under a null spatial model, providing a probabilistic assessment of how unlikely an observed cluster statistic is under the null hypothesis (Gustafson 1998). Waller and Jacquez (1995) formalized this approach by identifying five components of a spatial cluster test. The test statistic quantifies a relevant aspect of spatial pattern (e.g. Moran's  $I$ ). The alternative hypothesis describes the spatial pattern that the test is designed to detect. This may be a specific alternative, such as a circular cluster for the scan statistic, or it may be the omnibus "not the null hypothesis". The null hypothesis describes the spatial pattern expected when the alternative hypothesis is false (e.g. uniform cancer risk). The null spatial model is a mechanism for generating the reference distribution. This may be based on distribution theory, or it may use randomization (e.g. Monte Carlo) techniques. Most disease cluster tests employ heterogeneous Poisson and Bernoulli models for specifying null hypotheses (Lawson and Kulldorff, 1999). The reference distribution is the distribution of the test statistic when the null hypothesis is true. Comparison of the test statistic to the reference distribution allows calculation of the probability of observing that value of the test statistic under

the null hypothesis of no clustering. This five-component mechanism underpins most commonly used clustering methods.

There are dozens of cluster statistics (see Jacquez et al 1996a,b; Lawson and Kulldorff 1999, among others for reviews) that may be categorized for convenience as global, local, and focused tests. Global cluster statistics are sensitive to spatial clustering, or departures from the null hypothesis, that occur anywhere in the study area. Many early tests for spatial pattern, such as Moran's  $I$  (1950) are global tests. While global statistics can determine whether spatial structure (e.g. clustering, autocorrelation, uniformity) exists, they do not identify where the clusters are, nor do they quantify how spatial dependency varies from one place to another.

Local statistics such as Local Indicators of Spatial Autocorrelation LISA (Ord and Getis 1992) quantify spatial autocorrelation and clustering within the small areas that together comprise the study geography. Local statistics quantify spatial dependency (e.g. not significantly different from the null expectation, cluster of high values, cluster of low values, and high or low spatial outlier) in a given locality. Many local statistics have global counterparts that often are calculated as functions of local statistics. For example, Moran's  $I$  is the sum of the scaled local Moran statistics.

Focused statistics quantify clustering around a specific location or focus. These tests are particularly useful for exploring possible clusters of disease near potential sources of environmental pollutants. For example, Lawson (1989) and Waller et al. (1992) proposed tests that score each area for the difference between observed and expected disease counts, weighted by exposure to the focus (also see Lawson and Waller, 1996, for a review of these approaches). A commonly used exposure function is inverse distance to the focus ( $1/d$ ). The null hypothesis is no clustering relative to the focus, with expected number of cases calculated as the Poisson

expectation using the population at risk in each area and the assumption that risk is uniform over the study area.

Hundreds of cluster investigations are recorded in the literature, and several of these have resulted in cancer control activities such as epidemiological studies to understand potential causes. Notable examples of cluster studies include brain cancer (Kulldorff et al 1998), liver cancer (Zhan 2002), breast cancer (Roche et al 2002; Gregorio et al 2002), prostate cancer (Jemal et al 2002), colorectal cancer (Thomas and Carlin 2003), and cancer disparities (Hsu et al 2004), to name only a few. But to date and to our knowledge, none of these studies account for human mobility, and we thus do not know whether an observed cluster is real or explicable, in whole or in part, as an artifact of a clustering method that ignores residential histories.

How might we account for residential mobility in cluster studies? Hagerstrand (1970) conceptualized the *space time path* as an individual's continuous physical movement through space and time, and visually represented this as a 3-dimensional graph. Hornsby and Egenhofer (2002) recognized that space-time paths mediate individual-level exposure to pathogens and environmental toxins, and that practical application would require a mechanism for representing location uncertainty. A *space time prism* refers to the possible locations an individual could feasibly pass through in a specific time interval, given knowledge of their actual locations in the times bracketing that interval. The *potential path area* (Miller, 2004) shows the locations the individual could occupy given these constraints, and represents places where exposure events might occur. These constructs enabled new research approaches in diverse fields such as student life (Huisman and Forer, 1998), sports analysis (Moore et al, 2003), social systems (Kwan, 2000), transportation (Miller 1991), and the analysis of disparities in gender accessibility in households (Kwan 2003). While these approaches provide a proven mechanism for modeling

geospatial lifelines and related constructs, to date and to our knowledge there are no methods for the statistical evaluation of clustering among such lifelines other than the paper by Sinha and Mark (2005), who use Minkowski-type metrics to calculate a dissimilarity metric for geospatial lifelines, and then cluster this dissimilarity metric.

This paper proposes a novel technique for undertaking the statistical evaluation of clustering of residential histories for case-control data. We first develop the method, and then apply it to an ongoing case-control study of bladder cancer in southeastern Michigan.

## ***Methods***

We begin by defining an algebraic representation of residential histories, and a matrix representation that describes how spatial nearest neighbor relationships change through time. Next we develop a local case control cluster test, and then extend it to create global, local and focused tests at specific time points, and for entire residential histories. After completing development of the cluster statistics for residential histories, we next describe exposure traces that account for latency periods and exposure windows. We then develop clustering methods for exposure traces. After that we describe the bladder cancer data set that was analyzed with the new methods. In the Results section we describe application of the new cluster tests to evaluate possible clustering of residential histories of cases of bladder cancer in Michigan.

### **Cluster Methods for Residential Histories**

#### *Notation*

Define the coordinate  $\mathbf{u}_{i,t} = \{x_{i,t}, y_{i,t}\}$  to indicate the geographic location of the place of residence of the  $i^{\text{th}}$  case or control at time  $t$ . Residential histories for individual cases and controls can then be represented as the set of space-time locations as:

$$\mathbf{L}_i = (\mathbf{u}_{i0}, \mathbf{u}_{i1}, \dots, \mathbf{u}_{iT}) \quad (\text{Equation 1})$$

This defines individual  $i$  living at his or her place of residence found at  $\mathbf{u}_{i0}$  at the beginning of the study (time 0), and moving to location  $\mathbf{u}_{i1}$  at time  $t=1$ . At the end of the study individual  $i$  may be found at  $\mathbf{u}_{iT}$ .  $T$  is defined to be the number of unique observation times on all individuals in the study. This bears some emphasis as understanding of how  $T$  is recorded is essential in order to understand the cluster tests for residential histories. In other words,  $T$  is the total number of different observation times across all individuals, and so one might expect several locations in an individual residential history to be the same. For example, suppose we have 2 individuals (A and B) and record their residential histories. We record their places of residence at  $t=0$ , the beginning of the study. At some time  $t=1$  “A” moves to a different home, and moves again at time  $t=2$ . “B” never moves at all and hence has the location of the same initial place of residence recorded at times  $t=0, 1, \text{ and } 2$ . In this example  $T=2$ . Notice the duration between  $t=0$  to  $t=1$  may not equal the duration from  $t=1$  to  $t=2$ . This will be important later when we develop duration-weighted versions of the statistics.

While observations on residential histories occur at a finite number of time points or observation times, these observations do not have to happen at the same time for all individuals under scrutiny. When residential histories are self-reported, these observation times are defined by the “move” dates reported by the respondent. We modeled this as an instantaneous displacement from the spatial coordinates for entity  $i$  at time  $t$  ( $\mathbf{u}_{it}$ ) to those at time  $t+1$  ( $\mathbf{u}_{it+1}$ ). We defined this instantaneous displacement as occurring at time  $t+1$ . We viewed this as an observational model in which the entity is assumed to reside at its known location up until that moment when it is observed elsewhere (Figure 1).

Individual residential histories can be associated with time-dependent attributes such as weight, height, disease state, smoking status, case control status, and so on. These attributes may be associated with risk and thereby influence calculation of the latency period and exposure windows defined later. Later we also will use time of diagnosis to define exposure windows during which carcinogenesis was thought to have occurred. For now let us define a case-control identifier,  $c_i$  to be

$$c_i = \begin{cases} 1 & \text{if and only if } i \text{ is a case} \\ 0 & \text{otherwise} \end{cases} \quad (\text{Equation 2})$$

Define  $n_a$  to be the number of cases and  $n_b$  to be the number of controls. The total number of individuals in the study is then  $N=n_a+n_b$ .

### *k-Nearest Neighbor Relationships*

Let  $k$  indicate the number of nearest neighbors to consider when evaluating nearest neighbor relationships (see for example Jacquez 1996), and define a nearest neighbor indicator to be:

$$\eta_{i,j,k,t} = \begin{cases} 1 & \text{if and only if } j \text{ is a } k \text{ nearest neighbor of } i \text{ at time } t \\ 0 & \text{otherwise} \end{cases} \quad (\text{Equation 3})$$

We then can define a binary matrix of  $k^{\text{th}}$  nearest neighbor relationships at a given time  $t$  as:

$$\boldsymbol{\eta}_{k,t} = \begin{bmatrix} 0 & \eta_{1,2,k,t} & \cdot & \cdot & \eta_{1,N,k,t} \\ \eta_{2,1,k,t} & 0 & & & \cdot \\ \cdot & & \cdot & & \cdot \\ \cdot & & & \cdot & \eta_{N-1,N,k,t} \\ \eta_{N,1,k,t} & \cdot & \cdot & \eta_{N,N-1,k,t} & 0 \end{bmatrix} \quad (\text{Equation 4})$$

By convention we define  $\eta_{i,i,k,t} = 0$  (the diagonal elements) since we do not wish to count individuals as nearest neighbors of themselves. This matrix enumerates the  $k$  nearest neighbors (indicated by a 1) for each of the  $N$  individuals. The entries of this matrix are 1 (indicating that  $j$  is a  $k$  nearest neighbor of  $i$  at time  $t$ ) or 0 (indicating  $j$  is not a  $k$  nearest neighbor of  $i$  at time  $t$ ). It may be asymmetric about the 0 diagonal since nearest neighbor relationships are not necessarily reflexive. Since two individuals cannot occupy the same location<sup>1</sup>, we assume at any time  $t$  that any individual has  $k$  unique  $k$ -nearest neighbors. The row sums thus are equal to  $k$  ( $\eta_{i,\bullet,k,t} = k$ ) although the column sums vary depending on the spatial distribution of case control locations at time  $t$ . The sum of all the elements in the matrix is  $Nk$ . There exists a  $1 \times T+1$  vector of times denoting those instants in time when either (1) the system is observed and the locations of the entities are recorded, or (2) under continuous observation at least one entity changes geographic location. We can then consider the sequence of  $T$  nearest neighbor matrices defined by

$$\boldsymbol{\eta}_k^T = \{\boldsymbol{\eta}_{k,t}; t = 0..T\} \quad \text{(Equation 5)}$$

---

<sup>1</sup> While it is true that two individuals cannot occupy the exact same location (e.g. space occupied by one individual's body), residential history information can assign two individuals the same coordinate when they live in the same house. For two individuals with the same address this is not a problem since we would make them 1<sup>st</sup> order nearest neighbors of one another. It becomes a bit more complicated when 3 or more people occupy the same house, since we are uncertain as to how to assign nearest neighbor relationships. Two approaches have been proposed. The first creates fractional nearest neighbor weights (after Cuzick and Edwards 1990), the second propagates uncertainty in the nearest neighbor relationships by evaluating the permutations of possible nearest neighbors for the tied nearest neighbor relationships (Jacquez 1996). For the bladder cancer data presented later we have a case and a control that co-reside, and treat them as 1<sup>st</sup> order nearest neighbors of one another.

This defines the sequence of  $k$  nearest neighbor matrices for each unique temporal observation recorded in the data set, and thus quantifies how nearest neighbor relationships change through time. This demonstrates one way in which spatial weights (here the nearest neighbor relationship) can be specified from residential histories. We will now use these nearest neighbor relationships to construct case control spatial and space-time cluster tests for residential histories.

### *Spatially and Temporally Local Spatial Cluster Statistic*

A spatially and temporally local case-control cluster statistic is then:

$$Q_{i,k,t} = c_i \sum_{j=1}^N \eta_{i,j,k,t} c_j \quad (\text{Equation 6})$$

This is the count, at time  $t$ , of the number of  $k$  nearest neighbors of case  $i$  that are cases, and not controls (assuming  $i$  indeed is a case, if it isn't  $Q_{i,k,t} = 0$ ). Since a given individual  $i$  may have  $k$  unique nearest neighbors, this statistic is in the range  $0..k$ . It always is 0 when  $i$  is a control. When  $i$  is a case, low values indicate cluster avoidance (e.g. a case surrounded by controls), and large values (near  $k$ ) indicate a cluster of cases. When  $Q_{i,k,t} = k$ , all of the  $k$  nearest neighbors of case  $i$  are cases at time  $t$ .

### *Probabilities, Null Hypotheses and Randomization*

The statistical significance of  $Q_{i,k,t}$  may be evaluated using conditional randomization that holds the case control identifier for individual  $i$  fixed and then allocates the vector of remaining  $N-1$  case-control identifiers across the remaining individuals with a given probability function. If we assume equiprobability such that all individuals have equal disease risk we obtain:

$$P(c_j = 1 | c_i, H_{IV}) = \frac{n_a - c_i}{n_a + n_b - 1} \quad (\text{Equation 7})$$

Given the case-control identifier for individual  $i$ , this is the probability of individual  $j$  being a case under Goovaerts and Jacquez's (2004) neutral model Type IV ( $H_{IV}$ ) of spatial independence of risk for a spatially heterogeneous population density. As expressed in Equation 7, the exact number of cases ( $n_a$ ) and controls ( $n_b$ ) might not be reproduced under probabilistic sampling.

Their neutral model type V retains a specified level of spatial autocorrelation and may be simulated using rejection sampling, sequential indicator simulation, or conditional case-control index swapping to achieve the observed level of spatial autocorrelation. For imagery, Liebisch et al (2003) referred to this approach as conditional pixel swapping. Probabilities for neutral model type V are difficult to write in a closed form analogous to Equation 7.

Probabilities for neutral model type  $H_{VI}$  describe the situation where not all individuals have the same probability of being labeled a case. This occurs, for example, when we are concerned with detecting clusters that arise from additional risk *above and beyond* that of a background risk that is itself spatially heterogeneous. This may be accomplished in a variety of fashions to model known individual and environmental risk factors. Tests of the significance of  $Q_{i,k,t}$  are then identifying clusters of cases above and beyond that expected under the neutral model.

One calculates the value of the test statistic for each realization of the spatial distribution of cases generated under the chosen neutral model. These values under randomization are retained and used to construct the reference distribution of the statistic under the corresponding

null hypothesis. The observed value of the test statistic for the not randomized data (denoted  $Q_{ikt}^*$ ) is then compared to the reference distribution to calculate the probability:

$$P(Q_{ikt}^* | H_m) = \frac{(a+1)}{(b+1)} \quad (\text{Equation 8})$$

Here  $a$  is the number of conditional randomizations whose cluster statistic was greater than or equal to that observed for the not randomized data, and  $b$  is the total number of randomization runs conducted.

A convenient algorithm for conditional randomization under neutral model IV is to hold the case-control identifier for the  $i^{\text{th}}$  individual constant, and to then draw from the  $1 \times N-1$  vector of remaining case-control identifiers new case-control identifiers for the  $k$  nearest neighbors surrounding  $i$ . This sampling is accomplished without replacement. Alternatively, one could populate the  $k$ -nearest neighbors about  $i$  using the probabilities from Equation 7. This equation is correct for the first identifier so drawn, but needs to be adjusted for the second, third and so on. For the  $m^{\text{th}}$  identifier the correct probability for sampling without replacement is:

$$P(c_m = 1 | c_i, c_j \forall j = 1..m-1, H_{IV}) = \frac{n_a - c_i - \sum_{j=1}^{m-1} c_j}{n_a + n_b - m - 2} \quad (\text{Equation 9})$$

If one assumes sampling with replacement, so that the cases and controls are assumed drawn from a larger population, one can use Equation 7 without modification.

This approach does not work for neutral models type V and VI, since spatial structure in the background risk is lost. Instead one calculates the value of the test statistic for each of the  $N$  locations, for each realization of the spatial neutral model (of type V or VI) that produces a spatial point pattern of cases and controls with the desired level of spatial autocorrelation. The

probability assigned to clusters from these tests (as given by Equation 8) then accounts for the specified background variation in disease risk.

Note for each of the approaches listed above, that a reference distribution, test statistic, and corresponding p-value, may be calculated for each of the  $n_a$  case locations.

### *Simes Correction for local dependency*

The  $k$  P-values for the  $k$  individuals surrounding the  $i^{\text{th}}$  case are not independent of one another, as they necessarily will include one another as members of their own sets of  $k$  nearest neighbors. We therefore employ a modified Simes correction to account for the lack of spatial independence of the Q statistics. Such statistics are not independent because their case-control identifiers enter into calculation of the Q statistics for their neighbors. Their p-values should be adjusted for this lack of independence. We adjust the p-value using the Simes correction (Simes 1986), which is not as conservative as the Bonferroni correction. The Simes adjustment is calculated as  $p_i' = (k + 1 - a) p_i$ . Here  $k$  is the number of p-values being considered (the number of neighbors), and  $a$  is the index (starting at 1) indicating the location in the sorted vector of the p values for individual  $i$  and its neighbors.

### *Global Test for Spatial Clustering at Time $t$*

A global statistic for spatial clustering at time  $t$  may then be constructed as:

$$Q_{k,t} = \sum_{i=1}^N Q_{i,k,t} \quad (\text{Equation 10})$$

This is the space-time form of Cuzick and Edwards (1990) global test for case-control clustering. It is the count, over all cases, of the number of cases that are  $k$ -nearest neighbors to those cases at time  $t$ . One could divide this statistic, and others to follow, by  $n_a$  to facilitate the interpretation. For example, the test statistic would then be an average number of neighbor cases

per case instead of the integer total number of cases. This also would facilitate comparison across different studies. In this paper we will use the case-count version.

The probability of  $Q_{k,t}$  under  $H_{IV}$  is evaluated by allocating the case-control id's with equal probability over the  $N$  locations at time  $t$ .  $Q_{k,t}$  is then calculated and this process is repeated  $b$  times to construct the reference distribution and probability (Equation 8). Notice that since this is a global test conditional randomization that holds the case-control id for individual  $i$  constant is not needed.

### *Global Test for Spatial Clustering of Residential Histories*

A global test for spatial clustering among the  $N$  residential histories as represented in Equation 1 is

$$Q_k = \sum_{t=0}^T Q_{k,t} \quad (\text{Equation 11})$$

This is the sum, over all  $T+1$  time points, of the global statistic  $Q_{k,t}$ . It is a measure of the persistence of global clustering and is large when case clustering persists through time. Its reference distribution may be constructed under a randomization procedure in which the case-control ids are allocated with equal probability over the residential histories comprising the set

$$\{\mathbf{L}_i, i = 1..N\} \quad (\text{Equation 12})$$

### *Local Test for Spatial Clustering of Residential Histories through Time*

To determine whether cases tend to cluster through time around a specific case we may construct a test statistic:

$$Q_{i,k} = \sum_{t=0}^T Q_{i,k,t} \quad (\text{Equation 13})$$

For the  $i^{\text{th}}$  residential history, this is the sum, over all  $T+1$  time points, of the local spatial cluster statistic  $Q_{i,k,t}$ . It is the number of cases that are  $k$ -nearest neighbors of the  $i^{\text{th}}$  residential history (a case), summed over all  $T+1$  time points. It will be large when cases tend to cluster around the  $i^{\text{th}}$  case through time. Under neutral model type IV, the significance of  $Q_{i,k,t}$  is evaluated under a conditional randomization that holds the case id for  $i$  constant, and then allocates the remaining case-control id's at random over the  $N-1$  remaining residential histories. This statistic is useful for determining whether there is local clustering of residential histories about a specific case. The statistic can be calculated for all cases in the data set to identify those cases whose residential histories form local spatial clusters.

#### *Focused Test for Spatial Clustering at Time $t$*

Suppose one suspects the cases may be clustering about a specific focus defined by the lifeline (e.g. record of business addresses):

$$\mathbf{L}_F = \{\mathbf{u}_{F,0}, \mathbf{u}_{F,1}, \dots, \mathbf{u}_{F,T}\} \quad (\text{Equation 14})$$

This records the locations of the focus as it moves about through space-time, and includes instances in which the focus doesn't move as a degenerate instance as well as instances where the focus is mobile. A test for spatial clustering of cases about the focus at a given time  $t$  is then:

$$Q_{F,k,t} = \sum_{j=1}^N \eta_{F,j,k,t} c_j \quad (\text{Equation 15})$$

Here  $\eta_{F,j,k,t}$  is the nearest neighbor index indicating at time  $t$  whether the  $j^{\text{th}}$  individual is a  $k^{\text{th}}$  nearest neighbor of the geographic location of the focus defined by  $\mathbf{u}_{F,t}$ . The statistic  $Q_{F,k,t}$  is then the count of the number of  $k$ -nearest neighbors about the focus at time  $t$  that are cases. Under null hypothesis type IV randomization at time  $t$  may be accomplished by allocating the

case control identifiers with equal probability over the  $N$ -individuals. Since only the  $k$ -nearest neighbors are considered it is only necessary to allocate their indices. This may be accomplished by sampling without replacement from the  $1 \times N$  vector of the case-control identifiers, or by drawing the  $k$  required case control identifiers with probabilities defined by Equation 9 (for sampling without replacement) or Equation 7 (for sampling with replacement).

### *Focused Test for Spatial Clustering of Residential Histories about a Mobile Focus*

A test for focused clustering of residential histories through time is:

$$Q_{F,k} = \sum_{t=0}^T Q_{F,k,t} \quad (\text{Equation 16})$$

This is the count, over the  $T$  times, of the number of cases that are  $k$  nearest neighbors of the focus at each time point. This statistic is large when residential histories that are near the focus are cases. Its maximal value is

$$\max(Q_{F,k}) = kT . \quad (\text{Equation 17})$$

One drawback of using nearest neighbor relationships for focused tests is that the set of nearest neighbors to the focus are given equal weight in Equations 15 and 16, regardless of their actual geographic distance and direction with respect to the focus. But diffusion and active transport mechanisms that might carry emissions from the focus typically result in higher exposures near the focus, and it thus may make sense to use a maximum distance within which a set of  $k_i$  nearest neighbors are found. In these instances the set of nearest neighbors to the focus will vary (hence the  $i$  subscript denoting the  $i^{\text{th}}$  focus) depending on the number of cases and controls found within the specified distance of the focus.

### *Power of the Focused Tests and Specification of the Exposure*

Notice that the power of the tests given by Equations 15 and 16 decreases as  $k$  approaches  $N$  since  $Q_{F,k,t} = n_a$  when  $k=N$ , and its probability is then:

$$P(Q_{F,k,t} | H_0, k = N) = 1.0. \quad (\text{Equation 18})$$

When one wishes to search for clustering in instances where  $k$  approaches  $N$  power may be retained by constructing a weight function to model the hypothesized exposure. For geographically localized foci this may be based on proximity to the focus. One choice is

$$w_{F,j,t} = \frac{1}{r_{F,j,t}} \quad (\text{Equation 19})$$

Here  $r_{F,j,t}$  is the rank indicating proximity of the location of the  $j^{\text{th}}$  individual at time  $t$  (as given by  $\mathbf{u}_{j,t}$ ) to the location of the focus at time  $t$  ( $\mathbf{u}_{F,t}$ ). For example, the first nearest neighbor to the focus has rank 1, the second rank 2, and so on.

In many situations, such as airborne pollution or groundwater contamination, the magnitude of exposure is function not only of the proximity to the focus but also its orientation, since most dispersion processes (i.e. winds, infiltration through porous media) are anisotropic or direction-dependent. Depending on the amount of information available, exposure models of increasing complexity could be built.

An easy way to account for anisotropy is to replace the rank value  $r_{F,j,t}$  by a function of the separation vector  $\mathbf{h}_{jF,t} = |\mathbf{u}_{j,t} - \mathbf{u}_{F,t}|$  joining the location of the  $j^{\text{th}}$  individual at time  $t$  to the location of the focus at time  $t$ . Covariance functions seem to be natural choices for the weight functions  $w_{F,j,t}$  since it incorporate the spatial pattern of dependence of exposure data. For example, one could use the Exponential or Gaussian covariance functions defined as:

$$C_{Exp}(\mathbf{h}) = Exp\left(\frac{-3|\mathbf{h}|}{a(\theta)}\right) \quad (\text{Equation 20})$$

$$C_{Gaus}(\mathbf{h}) = Exp\left(\frac{-3|\mathbf{h}|^2}{[a(\theta)]^2}\right) \quad (\text{Equation 21})$$

where  $a(\theta)$  is the practical range of autocorrelation of the covariance models; that is the distance  $h$  at which the covariance function equals 0.05. This range is a function of the azimuth of the separation vector  $\mathbf{h}_{jF,t}$ . For example, the range of exposure to an airborne contaminant is expected to be larger in the direction of the prevailing winds.

More complex weight functions could be created if a process-based model of dispersion is available. For the example of airborne pollution, an atmospheric dispersion and deposition model could be developed to predict the fate of emissions and dust plumes from targeted facilities (Small et al., 1995). Such models require however many more parameters and assumptions concerning, for example, the emission rate, the meteorological conditions, complex terrain effects, the particle size and density for deposition calculation.

A limitation of process-based models is that they fail to provide a measure of uncertainty attached to their predictions and field exposure data are not readily incorporated. Geostatistics (Goovaerts, 1997) provides tools for modeling the spatio-temporal distributions of exposure and assessing the attached uncertainty. Various sources of information can be taken into account, such as measurements at a few monitoring stations, coordinates of major sources of exposure (i.e. factories) and transport characteristics (i.e. wind directions) that could be either directly incorporated into the prediction algorithm (Saito and Goovaerts, 2001) or fed into physical models to derive spatial trends (Goovaerts and Van Meirvenne, 2001). In the later case,

geostatistics is used to model the unexplained or residual part of the variability predicted by the process-based models.

The weight function, either based on geographic proximity (as in Equation 19) or derived using a process-based model or geostatistics (as in Equations 20 & 21), is then used to construct the weighted focused test at time  $t$  as:

$$Q'_{F,k,t} = \sum_{j=1}^N w_{F,j,t} c_j \eta_{F,j,k,t} \quad (\text{Equation 22})$$

The test for spatial clustering of residential histories about the focus through time is then:

$$Q'_{F,k} = \sum_{t=0}^T Q'_{F,k,t} \quad (\text{Equation 23})$$

Notice these weighted tests are conducted for the  $k$  nearest neighbors being considered.

When  $k=N$  the maximum values are:

$$\max(Q'_{F,k,t}) = \sum_{k=1}^N \frac{1}{k} \quad \text{and} \quad \max(Q'_{F,k}) = \sum_{t=0}^T \max(Q'_{F,k,t}) \quad (\text{Equation 24})$$

### *Duration-Weighted Tests for Clustering of Residential Histories*

The number of time points defined by the  $t=0, \dots, T$  observation times, and the frequency with which they are taken, can have some influence on the value of the above statistics. For example, many repeated observations when there is a chance of clustering could lead to spurious significance for the local and global tests for clustering of residential histories. We therefore develop duration-weighted versions of the tests. Define the duration of the  $t^{\text{th}}$  time period to be:

$$\omega_t = (t + 1) - t \quad (\text{Equation 25})$$

An observed statistic, such as the local case control cluster statistic  $Q_{i,k,t}$  has a given value for the period from  $t$  to  $t+1$ . At time  $t+1$  it takes on the new value defined by  $Q_{i,k,t+1}$ , and

that value pertains until the next observation time  $t+2$ . Notice the observation times are separated by periods that are not necessarily of equal duration, so  $\omega_t$  does not necessarily equal  $\omega_{t+1}$ . One then can define a duration weighted version of the local cluster statistic as

$$Q_{i,k,\omega_t} = Q_{i,k,t} \omega_t \quad (\text{Equation 26})$$

The units on the duration weighted version are case-time units (e.g. case days). Each of the statistics defined previously can now be redefined as duration weighted versions, with the proviso that all summations through time are from  $t=0$  to  $t=T-1$ . Specifically, the duration weighted version of the local statistic is:

$$Q_{i,k,w_t} = \sum \eta_{i,j,k,t} c_i c_j \omega_t \quad (\text{Equation 27})$$

The duration weighted global statistic is:

$$Q_{k,w_t} = \sum_{i=1}^N Q_{i,k,w_t} \quad (\text{Equation 28})$$

The duration weighted global statistic for clustering of residential histories is:

$$Q_k^\omega = \sum_{t=0}^{T-1} Q_{k,\omega_t} \quad (\text{Equation 29})$$

The duration weighted local statistic for clustering of residential histories is:

$$Q_{i,k}^\omega = \sum_{t=0}^{T-1} Q_{i,k,\omega_t} \quad (\text{Equation 30})$$

The duration weighted focused statistic over period  $\omega_t$  is:

$$Q_{F,k,\omega_t} = \sum_{j=1}^N \eta_{F,j,k,t} c_j \omega_t \quad (\text{Equation 31})$$

The duration weighted test for focused clustering of residential histories through time is:

$$Q_{F,k}^{\omega} = \sum_{t=0}^{T-1} Q_{F,k,\omega_t} \quad (\text{Equation 32})$$

The weighted focused test over duration  $\omega$  is:

$$Q'_{F,k,\omega_t} = \sum_{j=1}^N w_{F,j,t} c_j \eta_{F,j,k,t} \omega_t \quad (\text{Equation 33})$$

The weighted focused test, duration-based, for residential histories through time is:

$$Q'_{F,k}^{\omega_t} = \sum_{t=0}^{T-1} Q'_{F,k,\omega_t} \quad (\text{Equation 34})$$

When observations are made at regular time points such that  $\omega_0 = \omega_1 = \dots = \omega_{T-1}$  the not time weighted statistics may be used. When observations are recorded at irregular time intervals the duration-based statistics should be used. The not duration-weighted versions also can be used when one wishes to determine whether any of the  $T+1$  configurations of cases and controls are spatially clustered.

### *Accounting for Exposure Windows and Latency Periods*

When dealing with cancers, causative exposures may occur during an exposure window ( $\Delta_E$ ), followed by a latency period ( $\Delta_L$ ) before cancer is manifested and diagnosed. Given the residential history for case  $i$ ,  $\mathbf{L}_i$ , further denote the space-time coordinate representing place of residence at time of diagnosis as  $\mathbf{u}_{i,t_D}$ , noting that  $\mathbf{u}_{i,t_D} \in \mathbf{L}_i$ . We can then define that subset of the residential history  $\mathbf{L}_i$  over which the exposure window occurred as:

$$\mathbf{L}_i^E = \{\mathbf{u}_{i,t} \mid \forall (t_{i,D} - \Delta_L) > t > (t_{i,D} - \Delta_L - \Delta_E)\} \quad (\text{Equation 35})$$

Here  $t_{i,D}$  is the time of diagnosis for individual  $i$ . The term  $(t_{i,D} - \Delta_L)$  indicates the time prior to diagnosis when the latency period began and  $(t_{i,D} - \Delta_L - \Delta_E)$  is the time when the

causative exposure began. Hence equation 35 denotes that portion of individual  $i$ 's residential histories where causative exposures could have occurred. Notice that both the exposure window and latency period could be covariate-adjusted to account for risk factors such smoking and age. In this instance the latency period and exposure window vary from one individual to another and we write:

$$\mathbf{L}_i^E = \{\mathbf{u}_{i,t} \mid \forall (t_{i,D} - \Delta_{i,L}) > t > (t_{i,D} - \Delta_{i,L} - \Delta_{i,E})\} \quad (\text{Equation 36})$$

Here  $\Delta_{i,L}$  and  $\Delta_{i,E}$  are the latency period and exposure windows for the  $i^{\text{th}}$  individual. In either case (Equations 35 or 36) we call  $\mathbf{L}_i^E$  the *Exposure trace* for the  $i^{\text{th}}$  individual.

### *Clustering of Exposure Traces*

With the exposure trace defined we now can ask whether places of residence of individuals were spatially clustered while they were exposed, and whether the exposure traces themselves are spatially clustered. To do this we must first define the sampling distributions for the exposure traces, and then apply this sampling protocol to the controls.

Denote the distribution of exposure windows for the cases as  $\Psi_E$ . Notice this is a distribution of durations. This may be defined empirically as:

$$\hat{\Psi}_E = \{\Delta_{i,E}, i = 1, \dots, n_a\} \quad (\text{Equation 37})$$

Further, define the distribution of times of diagnosis as  $\Psi_D$ . This may be defined empirically as:

$$\hat{\Psi}_D = \{t_{i,D}, i = 1, \dots, n_a\} \quad (\text{Equation 38})$$

This is the distribution of points in time defined by the times of diagnosis of the cases. Finally, define the distribution of latency periods as  $\Psi_L$ . This may be defined empirically as:

$$\hat{\Psi}_L = \{\Delta_{i,L}, i = 1, \dots, n_a\}. \quad (\text{Equation 39})$$

### *Randomization Procedures for Exposure Traces*

In order to evaluate whether exposure traces of the cases cluster we first must construct a procedure for generating representative times of diagnosis, latency, and exposure periods under randomization. Once this is accomplished we will be able to determine whether the exposure traces for the cases cluster relative to those so constructed for the controls. Given the residential history of a case or control, the test proceeds as follows:

- (1) Specify a time of diagnosis by drawing from the distribution of times of diagnosis  $\hat{\Psi}_D$ .
- (2) Specify a latency period by drawing from the distribution of latency periods  $\hat{\Psi}_L$ .
- (3) Specify an exposure window by drawing from the distribution of exposure windows  $\hat{\Psi}_E$ .
- (4) From these, construct an exposure trace for that case control using Equation 36.
- (5) Undertake these 4 steps for all<sup>2</sup> of the cases and controls in the data set, resulting in candidate exposure traces for each.
- (6) Assign the case control identifiers across the residential histories with equiprobability (for neutral model type IV).

---

<sup>2</sup> With the exception of local versions of the statistics, for which one should hold the residential history and exposure trace of the individual being considered (“ego”) constant.

- (7) Calculate the desired test statistics for clustering of exposure traces, and repeat steps 1-6 a desired number of times to construct the reference distribution of the statistics under randomization.

The number of cases may differ from the number of controls and it thus makes sense to accomplish sampling under steps 1-3 with replacement in order to provide sample distributions arising as repeated, independent drawings from the parent distributions. This necessarily means the distributions obtained under sampling will not be exact replicates of the parent distributions. Notice this assignment procedure assumes time of diagnosis, and durations of the latency period and exposure windows are independent of one another. When this assumption isn't reasonable, the procedure would need to be modified to model such dependencies.

#### *Local Case-Control Test for the Spatial Clustering of Exposure Traces at Time $t$*

When health events such as cancers are caused by exposure to geographically localized factors we might expect the exposure traces for the cases to cluster relative to the exposure traces that are generated for the controls. The durations of the exposure traces may vary, and we therefore will employ duration-weighted statistics. We would like to know whether exposure traces for the cases exhibit spatial clustering relative to the controls both locally (to identify places where causative exposures occurred) and globally (to ascertain whether the exposure traces for the cases cluster when considered as a group). We also might wish to ask whether exposure traces for the cases exhibit focused clustering.

The exposure trace for case  $i$  ( $\mathbf{L}_i^E$ ) records those places where that individual lived during that time when exposures occurred that might have caused cancer later in life. Now define an indicator,  $e_{i,t}$ , as:

$$e_{i,t} = \begin{cases} 1 & \text{if and only if time } t \text{ is within the exposure trace for individual } i \\ 0 & \text{otherwise} \end{cases} \quad (\text{Equation 40})$$

When  $e_{i,t}$  is 1, let us say the exposure trace is “active”. A local case-control test for spatial clustering of exposure traces at time  $t$  is then:

$$Q_{i,k,t}^E = c_i e_{i,t} \sum_{j=1}^N \eta_{i,j,k,t} c_j e_{j,t} \quad (\text{Equation 41})$$

This is the count, at time  $t$ , of the number of  $k$  nearest neighbors of case  $i$ 's active exposure trace that are cases (and not controls) whose exposure traces also are active. Hence the statistic will be large when exposure traces of a group of cases are active at about the same time and cluster. Its value is 0 when individual  $i$  is a control, and also when individual  $i$  is a case with an inactive exposure trace. The duration weighted version of this statistic is:

$$Q_{i,k,\omega}^E = \omega_i c_i e_{i,t} \sum_{j=1}^N \eta_{i,j,k,t} c_j e_{j,t} \quad (\text{Equation 42})$$

### *Local Case-Control Test for the Spatial Clustering of Exposure Traces through Time*

We can explore whether active exposure traces of cases tend to cluster spatially through time. A statistic sensitive to this pattern is:

$$Q_{i,k}^E = \sum_{t=0}^T Q_{i,k,t}^E \quad (\text{Equation 43})$$

$Q_{i,k}^E$  will tend to be large when active exposure traces for cases tend to cluster around the active exposure trace of the  $i^{\text{th}}$  case. It will be 0 when  $i$  is a control, and small when a given case  $i$  has the traces of many controls as its neighbors. The duration-based version of this statistic is:

$$Q_{i,k}^{E,\omega} = \sum_{t=0}^{T-1} Q_{i,k,\omega}^E \quad (\text{Equation 44})$$

This statistic will be expressed in case-time units, indicating the number (for example) of case-days over the entire study period for which cases with active traces were  $k$ -nearest neighbors of the active trace of case  $i$ .

*Global Case-Control Test for the Spatial Clustering of Exposure Traces at Time  $t$*

We can ask whether, as a group, active case traces are spatially clustered relative to the active traces of the controls at a given time  $t$ . This is accomplished using the statistic:

$$Q_{k,t}^E = \sum_{i=1}^N Q_{i,k,t}^E \quad (\text{Equation 45})$$

This is simply the sum, over all cases, of the local statistic for clustering of case exposure traces at time  $t$ . This statistic will tend to be large when active traces of cases tend to be near one another, and small when the active traces of cases tend to have controls as their  $k$  nearest neighbors. The duration-based version is:

$$Q_{k,\omega_i}^E = \sum_{i=1}^N Q_{i,k,\omega_i}^E \quad (\text{Equation 46})$$

*Global Case-Control Test for the Spatial Clustering of Exposure Traces through Time*

A global test for the spatial clustering of the active exposure traces of cases through time is:

$$Q_k^E = \sum_{t=0}^T Q_{k,t}^E \quad (\text{Equation 47})$$

This is the sum, over all time periods, of the global cluster test for the clustering of exposure traces. It will be large when global clustering of active exposure traces tends to persist through time. The duration-based version of this statistic is:

$$Q_k^{E,\omega} = \sum_{t=0}^{T-1} Q_{k,\omega_t}^E \quad (\text{Equation 48})$$

### *Focused Case-Control Test for the Spatial Clustering of Exposure Traces at Time $t$*

We can also ask whether the exposure traces of cases cluster near putative emission sources. Again, these sources may be mobile, and we accomplish this by assigning larger weights for those cases that are near the focus. Recall from Equation 14 that we can represent a mobile source as  $\mathbf{L}_F = \{\mathbf{u}_{F,0}, \mathbf{u}_{F,1}, \dots, \mathbf{u}_{F,T}\}$ . The test for spatial clustering of cases about a focus at a given time  $t$  (Equation 15) may then be extended to be a focused test for clustering of exposure traces as:

$$Q_{F,k,t}^E = \sum_{j=1}^N \eta_{F,j,k,t} c_j e_{j,t} \quad (\text{Equation 49})$$

This is the count of the number of cases with active exposure traces that are  $k$  nearest neighbors of the focus at time  $t$ . Significance of this statistic may be evaluated by constructing exposure traces for the controls as described earlier, and by then repeatedly allocating case-control identifiers across the  $N$  lifelines that are  $k$  nearest neighbors of the focus in order to construct the reference distribution for  $Q_{F,k,t}^E$ . The duration weighted version of this statistic is

$$Q_{F,k,\omega_t}^E = \omega_t \sum_{j=1}^N \eta_{F,j,k,t} c_j e_{j,t} \quad (\text{Equation 50})$$

### *Focused Test for Spatial Clustering of Exposure Traces about a Mobile Focus through Time*

We can evaluate whether there is statistically significant clustering of exposure traces of cases about a mobile focus through time using the statistic:

$$Q_{F,k}^E = \sum_{t=0}^T Q_{F,k,t}^E \quad (\text{Equation 51})$$

This is the count, over  $T+1$  times, of the number of cases that have active exposure traces that are  $k$  nearest neighbors of the focus at each time point. The maximum value of this statistic is  $kT$ , and its significance may be evaluated under randomization by reallocating case-control identities over the exposure traces of the cases and controls as described in the previous section.

The duration-weighted version of this statistic is:

$$Q_{F,k}^{E,\omega} = \sum_{t=0}^{T-1} Q_{F,k,\omega_t}^E \quad (\text{Equation 52})$$

### *Weighted Focused tests for Exposure Traces*

The power of the k-nearest neighbor based focused test for exposure traces decreases as  $k$  approaches  $N$ . Weights such as that suggested in Equations 19-21 may be used to construct a weighted focused test for exposure traces at a given time  $t$ :

$$Q_{F,k,t}^{\prime E} = \sum_{j=1}^N w_{F,j,t} \eta_{F,j,k,t} c_j e_j \quad (\text{Equation 53})$$

The test for focused clustering of exposure traces through time is then:

$$Q_{F,k}^{\prime E} = \sum_{t=0}^T Q_{F,k,t}^{\prime E} \quad (\text{Equation 54})$$

The significance of these statistics is evaluated using randomization across the  $k$  nearest neighbors of the focus as described earlier. The corresponding duration-weighted versions are

$$Q_{F,k,\omega_t}^{\prime E} = \omega_t \sum_{j=1}^N w_{F,j,t} \eta_{F,j,k,t} c_j e_j \quad (\text{Equation 55})$$

This is the weighted focused test over duration  $\omega_t$ . The duration-based weighted focused test for exposure traces through time is

$$Q_{F,k}^{E\omega} = \sum_{t=0}^T Q_{F,k,\omega_t}^E \quad (\text{Equation 56})$$

## Bladder Cancer in southeastern Michigan

A population-based bladder cancer case-control study is currently underway in southeastern Michigan. Cases are recruited from the Michigan State Cancer Registry and diagnosed in the years 2000-2004. Controls are frequency matched to cases by age ( $\pm 5$  years), race, and gender, and recruited using a random digit dialing procedure from an age-weighted list. To be eligible for inclusion in the study, participants must have lived in the eleven county study area for at least the past 5 years and had no prior history of cancer (with the exception of non-melanoma skin cancer). Participants are offered a modest financial incentive and research is approved by the University of Michigan IRB-Health Committee.

The data presented here are from 63 cases and 182 controls. As part of the study, participants complete a written questionnaire describing their residential mobility history. The duration of residence and exact street address were obtained, otherwise the closest cross streets were provided. Each residence in the study area was geocoded and assigned a geographic coordinate in ArcGIS; residences outside the study area were not geocoded. Participants resided at 1004 homes within the study area, with time spent averaging 64% of their lifetimes. Residences within the study area were successfully geocoded: 76% automatically matched using ArcGIS settings of spelling sensitivity equal to 75, minimum candidate score equal to 10, and a minimum match score equal to 60. The unmatched addresses were manually matched using cross streets with the assistance of internet mapping services (15%). If cross streets were not provided, best informed guess placed the address on the road (5%), and as a last resort, residence was matched to town centroid (4%). At the time of this writing geocoding and data collection

are ongoing, hence the results reported in this manuscript are entirely preliminary and should not be used to draw any conclusions regarding the spatial patterns of bladder cancer in Michigan. The analysis undertaken in the manuscript is provided only as an example application of the new  $Q$  statistics.

Industrial histories have also been collected for the study area, and will be explored to explain local clustering. Industries reported to or believed to emit contaminants that have been associated with bladder cancer were identified using the Toxics Release Inventory (USEPA 2000) and the Directory of Michigan Manufacturers (Manufacturer Publishing Co., 1946, 1953, 1960, 1969, 1977, 1982). Standard Industrial Classification (SIC) codes were adopted, but prior to SIC coding, industrial classification titles were selected. Characteristics of 268 industries, including, but not limited to, fabric finishing, wood preserving, pulp mills, industrial organic chemical manufacturing, and paint, rubber, and leather manufacturing, were compiled into a database. Industries were geocoded following the same matching procedure as described for residences: 89% matched to the address, 5% were placed on the road using best informed guess, and as a last resort, 6% were matched to town centroid. Each industry was assigned a start year and end year, based on best available data. The data on these industries is used to demonstrate the focused versions of the  $Q$  statistics.

## **Results**

To demonstrate the methods we implemented the local and global  $Q$  statistics for clustering of residential histories, specifically the local test at time  $t$ ,  $Q_{i,k,t}$  (Equations 6), and its global counterpart  $Q_{k,t}$  (Equation 10). We also implemented the local test for clustering of residential histories through time  $Q_{i,k}$  (Equation 13), and the global test for clustering of residential histories  $Q_{i,k}$  (Equation 11). We also were concerned with possible clustering of

cases near the industrial facilities, and evaluated this using the focused test at time  $t$   $Q_{F,k,t}$  (Equation 15) as well as the focused test through time  $Q_{F,k}$  (Equation 16). In addition we programmed the duration-weighted versions of these statistics, and for the focused tests we also employed exposure weights calculated using the inverse rank distance (Equation 19).

*Results for  $Q_{kt}$*  These techniques were implemented in TerraSeer's STIS software using the Application Programmer's Interface. This allowed us to create a methods dynamic linked library with our new techniques that we then invoked using an automatically generated dialog. Time animated maps of the places of residence of the cases and controls, and of the changing geography of the municipal water supplies, were constructed using STIS (Figure 2). These display the changing geography of the cases and controls as they move from one place to another, alterations in the geography of the municipal water supplies as they are founded, expand and merge, as well as township boundaries. To verify the methods we compared results using the  $Q$  statistics to those obtained using Cuzick and Edward's test in the ClusterSeer software. Specifically, we used STIS to calculate the  $Q_{kt}$  statistics through time and then exported the data for July 1, 1969. We choose this time point because  $Q_{kt}$  reached a local peak of  $Q_{kt}=77$  that was statistically significant (see Figure 3). The Cuzick and Edward's test in ClusterSeer returned  $T_5=77$ , confirming the results from STIS. As noted earlier, Cuzick and Edward's test is a special case of the  $Q$ -statistic for the global test at time  $t$ ,  $Q_{kt}$ . Note that  $Q_{kt}$  is calculated as the sum of the local  $Q$  statistics at time  $t$ ,  $Q_{ikt}$ , and thereby provides verification that the statistic  $Q_{ikt}$ , from which the family of  $Q$  statistics is derived, is being calculated correctly. The graph of  $Q_{kt}$  through time (Figure 3) is ascending, reflecting the larger number of cases in the latter time periods. We found five periods when cases were significantly clustered relative to the controls: January 1 1929 through January 1 1935, January 1 1941 through November 26 1942, January 1

1960 through January 1 1961, August 22 1967 through January 1 1975 and January 1 1995 through January 1 1997. We must remind the reader that these results are highly preliminary and that data collection is incomplete. In fact, and as noted later in the Discussion, it is likely the observed clustering in these data is due to the geographic ordering in which the data are being collected. Nonetheless, this example demonstrates how plots of the  $Q_{kt}$  statistics may be used to evaluate case clustering of residential histories through time.

*Results for  $Q_{k,w_t}$*  The results reported above were not time standardized. We therefore undertook an analysis using the time-standardized version of  $Q_{kt}$  called  $Q_{k,w_t}$  as per Equation 28. This expresses the amount of clustering at a given time interval in cases per unit time period. STIS reports times down to the second, hence results are recorded in person seconds. Figure 4 shows a similar overall increasing trend but also a greater variability in the value of the Q statistic through time. This is driven both by the increased number of cases through time and also by differences in the durations between movement events. When these sources of variability are accounted for we find episodic case clustering in approximately the same time intervals as found for the not time weighted statistic.

*$Q_{i,k}$  to evaluate Clustering of Residential Histories* The statistics  $Q_{kt}$  and  $Q_{k,w_t}$  are sensitive to a clustering of cases relative to the controls, and are evaluated at each of the T+1 time points in the set of residential histories. We also can ask, whether residential histories of the cases cluster near the residential histories of other cases by using the statistics  $Q_{i,k}$  (Equation 13) and its duration-weighted version  $Q_{i,k}^{\omega}$  (Equation 30). Since our analysis above demonstrated the results are not overly sensitive to duration weighting, we report results only for the not-weighted tests. This test will associate a statistic and a p-value with each residential history. The distribution of the statistic versus its p-value is shown in Figure 5. A map of the residential

histories on April 12, 1997 is shown in Figure 6. Note the two red dots that denote the place of residence of the two cases with statistically significant clustering of residential histories. Over the entire time span of the study, these two cases tend to be surrounded by residential histories of other cases, rather than the residential histories of controls. Because of residential mobility, the two red dots move about through time. This animation is quite compelling in the STIS and is approximated by the simpler animation in Figure 2. Note the animation in Figure 2 is sampled from the complete animation created when running the STIS software. This is necessary to create .avi files of small enough size for effective posting on the internet. Periods in which a red dot disappears from the animation denote time periods when that individual moved out of the study area. To summarize, there is statistically significant clustering of the residential histories of cases about the two cases shown in red in Figure 2.

*Focused clustering* To demonstrate the use of the focused versions of the  $Q$  statistic we analyzed possible clustering of the residential histories of cases near the 268 industrial facilities that produced compounds thought to be putative carcinogens for bladder cancer. We undertook two sets of analyses using  $Q_{F,k}$  (Equation 16). The first evaluated focused clustering of residential histories using the full set of  $k=5$  nearest neighbors. The second only considered those nearest neighbors within 1 kilometer of the focus.

When considering the 5 nearest neighbors to each industry, 24 of the 268 industrial facilities had p-values less than 0.05 (Figure 7). Thus under the null hypothesis that each person in the study had an equal probability of being labeled a case, these 24 candidate foci had a significant excess of cases among each of their five nearest neighbors, at least at the nominal 0.05 level. Notice that at the 0.05 level, we would have expected 13.4 foci to be significant under this null hypothesis. Using an experiment-wise error approach, and a 5% critical value,

the adjusted alpha level of the test is 0.000187 using the Bonferonni correction, and is 0.000191 using Sidak's multiplicative inequality. Using 49,999 randomizations, we were able to resolve p-values as small as 0.00005. None of these industries proved to be statistically significant foci once multiple testing was accounted for.

We also used the distance-based approach considering those neighbors within 4,000 m of each industrial facility. Under this approach, 10 industrial facilities had p-values  $< 0.05$  (Figure 8), but none of these were significant once multiple testing was accounted for.

## ***Discussion***

This paper presented an entirely new approach to evaluating case control clustering of residential histories. To date and to our knowledge, almost all case control cluster tests rely on the static view, analyzing clustering at one point in time or independently at several points in time. By using the mathematical construct of a residential history in Equation 1, and the notion of super sets of proximity matrices (Equation 5) to represent the changing geometry of place of residence, we have derived local, global and focused tests that are realistic in the sense that they quantify human residential mobility.

The results of the analysis of the bladder cancer data are entirely preliminary, and should not be interpreted to reach any inferences or conclusions regarding case-control clustering of bladder cancer in Michigan. At the time of this writing we believe statistically significant spatial clustering of cases is the result of a geographic pattern in the temporal order in which cases are reported. Because of recent implementation of the HIPPA (Health Insurance Portability and Accountability Act) legislation, The University of Michigan hospital systems has been unwilling to release case data until its official position on these requirements was completely formulated. As a result, bladder cancer cases that were treated at the University of

Michigan hospitals are only now being recruited to the study's data set. Because many of these cases are from the surrounding environs of Washtenaw and Livingston counties, the data set analyzed in this paper has a deficit of cases in these areas. Selection of controls employs a population sample using random digit dialing, and appropriately represents the entire study area. As a result, there is a statistically significant deficit of cases in Washtenaw and Livingston counties, and a concomitant clustering of cases in the balance of the study area. We intend to revisit this analysis once the data set is complete.

The approaches detailed in this paper are general in the sense that other weight structures such as inverse distance and adjacency could be used in place of  $k$ -nearest neighbor relationships. We choose to work with nearest neighbor measures because of their acknowledged superiority to adjacency- and distance-based measures (see for example Jacquez 1996).

We could not demonstrate each of the statistics developed in this paper, due to both data and space constraints. We note that exposure traces could be implemented to represent cases and controls of similar ages, in addition to those at a point in time. For example, a researcher may wish to determine whether cases cluster together when they were children, irrespective of year, thereby indicating early-lifetime vulnerability to an environmental exposure in the area. These clustering tools thus can be used to display cancer clusters of similarly-aged participants, as well as clusters based on the years a participant lived at a residence. In this manner, clusters of children can be investigated, whether they are born in the same generation or born in different generations.

In conclusion, the methods presented in this paper account for residential mobility and are thus far more realistic than existing tests that are founded on static geographic representations. They

thus are preferred over clustering methods that ignore human mobility. The methods demonstrated in this paper have been programmed in a dynamic linked library that can be obtained from the first author and used in conjunction with a STIS.

### ***Competing Interests***

Geoffrey Jacquez is President of BioMedware, the software company that is developing the STIS software.

### ***Author's Contributions***

Geoffrey Jacquez derived the statistics and drafted the majority of this manuscript. Andy Kaufmann programmed and tested the statistical methods in the STIS software. Gillian AvRuskin and Jaymie Meliker provided data and drafted the description of the data sets. Pierre Goovaerts wrote the sections on geostatistical weighting functions for the focused tests. Jerome Nriagu is Principal Investigator on the R01 project that is collecting the bladder cancer data set.

### ***Acknowledgements***

This study was supported by grant R01 CA96002-10, Geographic-Based Research in Cancer Control and Epidemiology, from the National Cancer Institute. Development of the STIS<sup>TM</sup> software was funded by grants R43 ES10220 from the National Institutes of Environmental Health Sciences and R01 CA92669 from the National Cancer Institute. Access to cancer case records was provided by Michigan Cancer Surveillance Program within the Division for Vital Records and Health Statistics, Michigan Department of Community Health. The authors thank Michigan Public Health Institute for conducting the telephone interview and Stacey Fedewa and Lisa Bailey for entering written surveys into a database. Thanks to Wanda

Angelomatis and Fred Wallace who hosted the first author and his daughter for two weeks at Berkenhead Lake in British Columbia where these methods were originally formulated.

## **References**

- Collia DV, Sharp J, Giesbrecht L. The 2001 National Household Travel Survey: a look into the travel patterns of older Americans. *J Safety Research*. 34(4):461-70, 2003.
- Cuzick, J., and R. Edwards. 1990. Spatial clustering for inhomogeneous populations. *Journal of the Royal Statistical Society Series B* 52: 73-104.
- Goodchild, M. (2000). GIS and Transportation: Status and Challenges. *GeoInformatica* 4: 127-139.
- Goovaerts, P. 1997. *Geostatistics for Natural Resources Evaluation*. Oxford University Press, New York.
- Goovaerts, P. and M. Van Meirvenne. 2001. Delineation of hazardous areas and additional sampling strategy in presence of a location-specific threshold. In P. Monestiez et al., editors, *geoENV III - Geostatistics for Environmental Applications*, pages 125-136. Kluwer Academic Publishers, Dordrecht.
- Goovaerts, P and G. Jacquez. 2004. Accounting for regional background and population size in the detection of spatial clusters and outliers using geostatistical filtering and spatial neutral models: the case of lung cancer in Long Island, New York. *International Journal of Health Geographics* 2004, 3:14.
- Gregorio DI, Kulldorff M, Barry L, Samociuk H. 2002. "Geographic differences in invasive and in situ breast cancer incidence according to precise geographic coordinates, Connecticut, 1991-95." *Int J Cancer* 100:194-8.

- Greiling D. A., G. M. Jacquez, A. M. Kaufmann, and R. G. Rommel (2005). "Space time visualization and analysis in the Cancer Atlas Viewer". *Journal of Geographical Systems* (Accepted).
- Gustafson, E. J. 1998. "Quantifying landscape spatial pattern: What is the state of the art?" *Ecosystems*(1): 143-156.
- Hagerstrand, T, 1970. What about people in regional science? *Papers of the Regional Science Association*, 24: 7-21.
- Hornsby, K and M. Egenhofer 2002. Modeling moving objects over multiple granularities, Special issue on Spatial and Temporal Granularity, *Annals of Mathematics and Artificial Intelligence*. 36: 177-194
- Hsu C E, H E Jacobson and F S Mas. 2004. "Evaluating the disparity of female breast cancer mortality among racial groups; a spatiotemporal analysis" *International Journal of Health Geographics* 3:4
- Huisman, O., and P. Forer. 1998. "Computational agents and urban life spaces: A preliminary realization of the time-geography of student lifestyles". *Proceedings of the 3rd International Conference on GeoComputation*. University of Bristol, UK, Sept. 1998.
- Jacquez, G.M. 1996. "A k-nearest neighbor test for space-time interaction". *Statistics in Medicine*. 15:1934-1949.
- Jacquez, G.M., R. Grimson, L. Waller and D. Wartenberg. 1996. 'The analysis of disease clusters Part 2: Introduction to techniques'. *Infection Control and Hospital Epidemiology*, 17:385-397.
- Jacquez, G.M., L. Waller, R. Grimson and D. Wartenberg. 1996. 'The analysis of disease clusters Part I: State of the art'. *Infection Control and Hospital Epidemiology*, 17:319-327

- Jacquez, G. M., D. Greiling and A. Kaufmann. 2005. "Design and Implementation of Space Time Information Systems". *Journal of Geographical Systems*. (Accepted).
- Jemal A, Kulldorff M, Devesa SS, Hayes RB, Fraumeni JF. 2002. "A geographic analysis of prostate cancer mortality in the United States." *International Journal of Cancer* 101:168-174.
- Klepeis, NE, Nelson WC, Ott WR, Robinson JP, Tsang AM, Switzer P, Behar JV, Hern SC, Engelmann WH. The National Human Activity Pattern Survey (NHAPS): a resource for assessing exposure to environmental pollutants. *J Exposure Analysis and Environmental Epidemiology* 11:231-252, 2001.
- Kulldorff, M., Wathas, W. F. et al. 1998. "Evaluation of cluster alarms: A space-time scan statistic and brain cancer in Los Alamos". *American Journal of Public Health*. 88: 1377-1379.
- Kwan, M.P. et al 2003. Accessibility in space and time: A theme in spatially integrated social science. *Journal of Geographical Systems*. 5:1-3.
- Lawson, A.B. 1989. "Score tests for detection of spatial trend in morbidity data." Dundee: Dundee Institute of Technology.
- Lawson AB, Kulldorff M. 1999. "A review of cluster detection methods." In *Advanced Methods of Disease Mapping and Risk Assessment for Public Health Decision Making*, Lawson et al. (eds). London: Wiley, 99-110.
- Lawson, A. B. and L. A. Waller. 1996. "A review of point pattern methods for spatial modelling of events around sources of pollution." *Environmetrics* 7: 471-487.

- Liebisch, N, G. M. Jacquez, P. Goovaerts and A. Kaufmann. 2002. "New methods to generate neutral images for spatial pattern recognition", Lecture Notes in Computer Science, 2478: 181-195, Springer-Verlag Berlin Heidelberg.
- Mather, F J, L E Whited, E. C Langlois, C F Shorter, C M. Swalm, J G. Shaffer and WR Harley. 2004. Statistical methods for linking health, exposure and hazards. Environmental Health Perspectives 112:1440-1445.
- Meliker, J, M. Slotnick, GA AvRuskin, A Kaufmann, GM Jacquez and JO Nriagu. 2005. Improving exposure assessment in environmental epidemiology: applications of a Space-Time Information System. Journal of Geographical Systems (Accepted).
- Miller, H. J. 1991. "Modeling accessibility using space-time prism concepts within geographical information systems," International Journal of Geographical Information Systems, 5, 287-301.
- Miller, H. 2004. A measurement theory for time geography. Geographical Analysis. (In Press)
- Moore, A.B. et al 2003. Proceedings of the 7th International Conference on GeoComputation University of Southampton, United Kingdom 8 - 10 September 2003.
- Moran, P.A.P. 1950. "Notes on continuous stochastic phenomena." Biometrika 37: 17-23.
- Ord J.K. and Getis A. 1995. Local spatial autocorrelation Statistics: Distributional issues and an application. Geographical Analysis 27:286-306.
- Reuscher TR, Schmoyer RL, Hu PS. Transferability of Nationwide Personal Transportation Survey data to regional and local scales. Transport Res Rec (1817): 25-32, 2002.
- Roche LM, Skinner R, Weinstein RB. 2002. "Use of a geographic information system to identify and characterize areas with high proportions of distant stage breast cancer." J Public Health Manag Pract 8:26-32.

- Simes, R.J. 1986. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 73: 751-4.
- Sinha, G. and D. M. Mark. 2005. Measuring Similarity between Geospatial Lifelines in Studies of Environmental Health. *Journal of Geographical Systems* (Accepted).
- Saito, H. and P. Goovaerts. 2001. "Accounting for source location and transport direction into geostatistical prediction of contaminants." *Environ. Sci. Tech.* 35: 3924-3930.
- Small MJ, Nunn AR, Forslund BL, Daily DA. 1995. "Source attribution of elevated residential soil lead near a battery recycling site." *Environ. Sci. Tech.* 29: 883-895.
- Steinmaus, C., Y. Yuan, M. N. Bates, and A. H. Smith (2003), Case-control study of bladder cancer and drinking water arsenic in the Western United States, *Am. J. Epidemiol.*, 158, 1193-1201.
- Thomas A., Carlin B.P. 2003. "Late detection of breast and colorectal cancer in Minnesota counties: An application of spatial smoothing and clustering." *Statistics in Medicine* 22:113-127.
- Waller, L.A., B.W. Turnbull, L.C. Clark, and P. Nasca. 1992. Chronic disease surveillance and testing of clustering of disease and exposure: Application to leukemia incidence and TCE-contaminated dumpsites in upstate New York. *Environmetrics* 3: 281-300.
- Waller, L. A. and G. M. Jacquez. 1995. "Disease models implicit in statistical tests of disease clustering." *Epidemiology* 6(6): 584-590.
- Pickle, LW. LA Waller, and AB Lawson. 2004. "Current practices in cancer spatial data analysis: a call for guidance" *International Journal of Health Geogrphics* (In press).
- Zhan FB. 2002. "Are deaths from liver cancer, kidney cancer, and leukemia clustered in San Antonio?" *Texas Medicine* 98:51-6.



Figure 1. 3-d graphical representation of residential histories from Equation 2 using the instantaneous displacement movement model. Individual  $i$  moves from location  $\mathbf{u}_{i,0}$  to  $\mathbf{u}_{i,1}$  at time  $t=1$ , and stays at that place of residence until  $t=T$ . Individual  $j$  stays at the same place of residence from  $t=0$  to  $t=T$ .

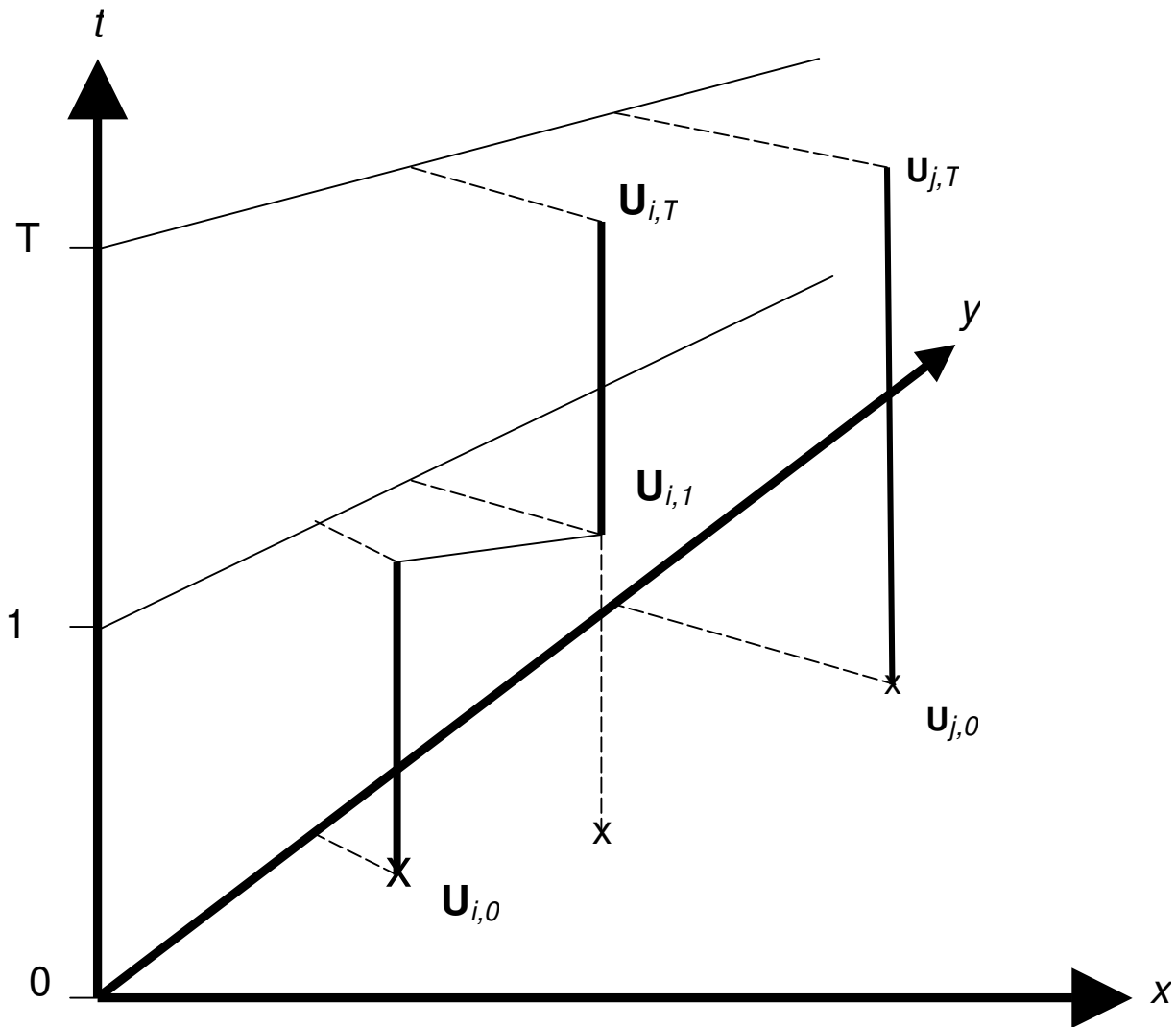


Figure 2. Residential histories of cases and controls in southeastern Michigan. Cases are shown as circles, controls as squares. The residential histories of cases that are adjacent to other case residential histories are shown in red. Click on the icon to start the animation.



C:\Documents and  
Settings\Jacquez\Des

Figure 3. Graph of  $Q_{tk}$  (top) and its Probability (bottom) through time for  $k=5$ .  $Q_{tki}$  is the count of the number of  $k^{\text{th}}$  nearest neighbors of case  $i$  that also are cases, and  $Q_{tk}$  is the sum of over all cases of the  $Q_{tki}$ . Shown in red are those time intervals in which the probability of  $Q_{tk}$  was 0.01 or smaller. The significance of  $Q_t$  is obtained under conditional randomization by generating a shuffled list of case control identifiers, and then for each individual replacing the case control identifier from the shuffled list with the observed value when its contribution to the global  $Q_t$  is considered.

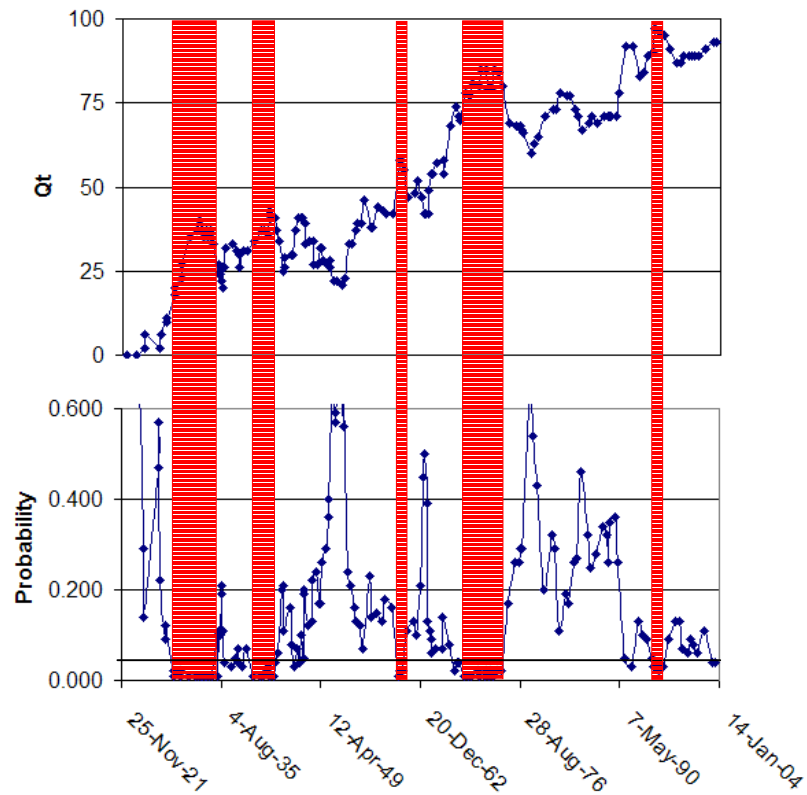


Figure 3

Figure 4. Graph of  $Q_{k,w_t}$  (top) and its Probability (bottom) through time for  $k=5$ .  $Q_{k,w_t}$  is the time weighted version of  $Q_{tki}$  and is expressed in case-seconds.

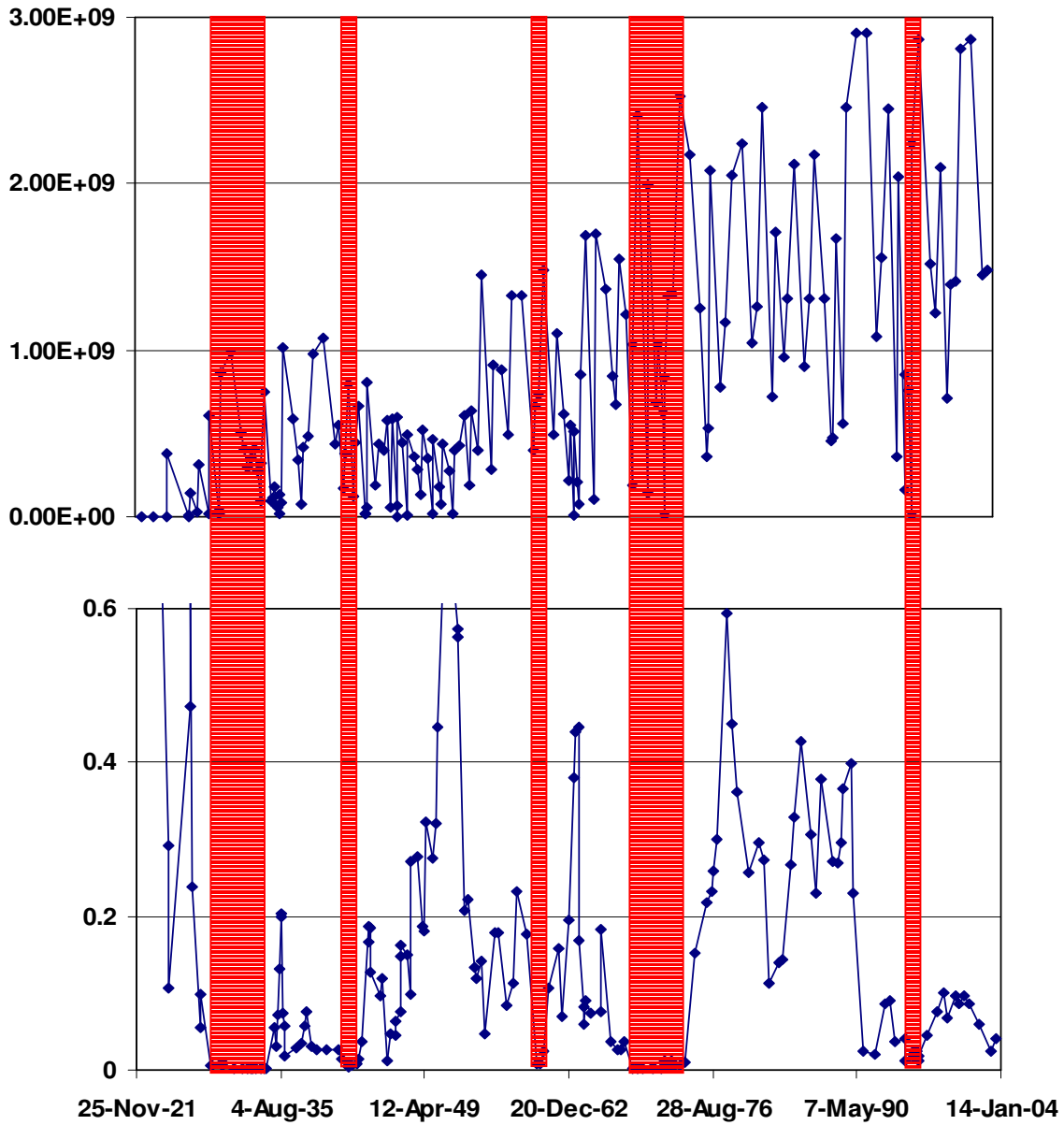


Figure 5. Plot of the probability of  $Q_i$  as a function of  $Q_i$ . Each point in the scatterplot represents the residential history of a case. The two residential histories with p-values less than 0.05 tend to be near the residential histories of other cases to a statistically significant extent.

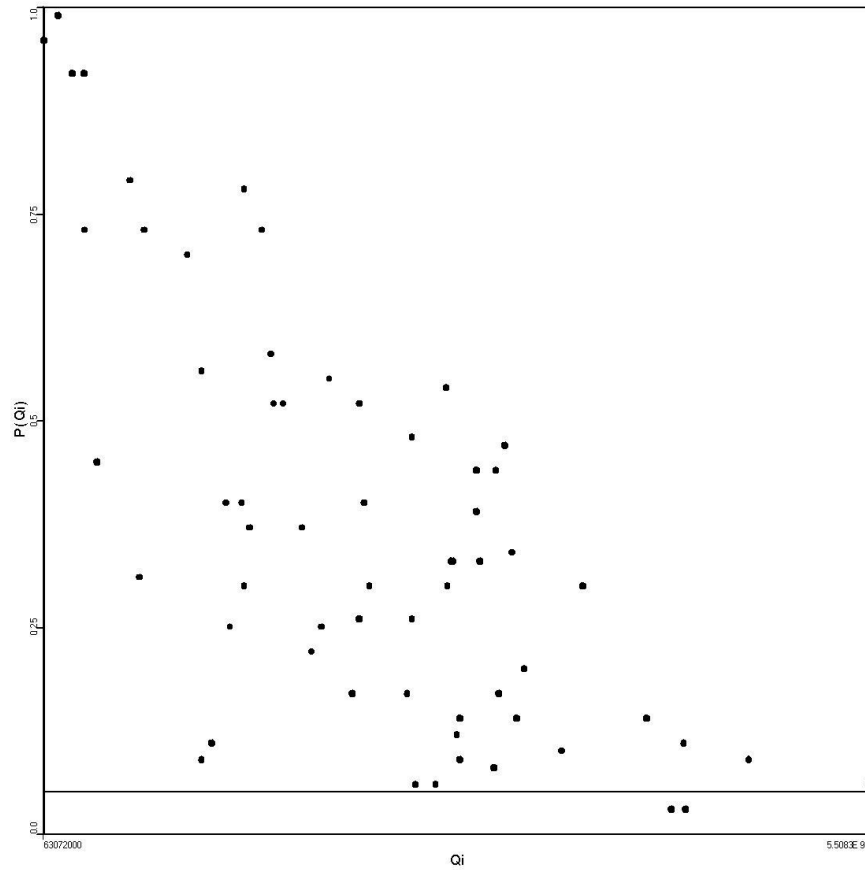


Figure 6. Map of cases and controls on 4/12/1997. Cases are shown as dots within a circle, controls are shown as crosses. The two cases whose residential histories tend to be surrounded by the residential histories of other cases are shown in red.

