A Comparison of two Significance Testing Methodologies or the Knox Test

Theophilides C. N.[1], Binkowski E. S.[2], Ahearn S. C.[1] and Paul W. S.[3]
[1]Center for Advanced Research of Spatial Information, Hunter College of The City University of New York New York, USA, E-mail: sahearn@hunter.cuny.edu
[2]Department of Mathematics and Statistics, Hunter College of The City University of New York,USA
[3]Epidemiology and Disease Control Programs, Chicago Department of Health, USA

Abstract
The Knox method for detecting space-time interactions of point data has been widely used in infectious Epidemiology, and crime mapping. However, the parametric methods (chi-square, Poisson) used for testing the significance of space-time interaction suffer from a major flaw, the violation of the assumption of independence of pairs. A Monte Carlo method that consists of random switching of the time labels of points is also flawed because in the case of heavy clustering of either dimension (space or time) the power of the test is reduced due to the switching of already close labels. Recently, a new Monte Carlo method for testing significance has been proposed which consists of completely random sampling in space and time, i.e., an unconditional extension of the Knox test. Comparative evaluation of these statistical tests and empirical methodologies has never been conducted. Here we present the first comparative examination between the chi-square test and the random Monte Carlo unconditional extension of the Knox test. We use a space-time version of the kappa statistic for evaluation and show that the results of the unconditional Monte Carlo methodology are significantly different and superior to those of the chi-square test.

1. Introduction

The occurrence of human disease is often presented as point data. Those points can also be labeled with time (or date) of appearance. It is often of interest whether there is clustering in either the space and/or time dimension for these points. Geographers usually examine the spatial clustering component (Heining, 2003) and epidemiologists are more interested in the temporal sequence of occurrence of the disease events (Lawson, 2001). Clustering is often only evident if both dimensions are examined simultaneously (Knox, 1963). Knox (1963, 1964, 1964b) introduced a way of detecting spatial and temporal interaction between disease incidents and inferring an infectious process from this space-time interaction (Andersson et al., 1995 and Alexander, 1992). The Knox test consists of creating all possible pair combinations of data points found in a spatial and temporal domain and then examining the interpoint distances in space and time and classifying the pairs as close or not in either space or time. The total number of pairs found close in space and time is compared to what would be expected by chance given the individual number of close space and close time pairs. Knox (1963) initially proposed a chi-square test for evaluating the significance of the interaction and then a Poisson test (1964). Both of these tests require that the data (pairs) being evaluated are part of an independent process. Because the pairs are formed by a finite number of points and they share points, this requirement is violated (Knox, 1964). Moreover, excess clustering in either the space or time dimension of the data can exacerbate this violation, by resulting in underestimation of the variance. Mantel (1967) proposed a solution to the significance test that involved the Monte Carlo switching of the space-time labels of the data points and then ranking the actual number of close space-time pairs to the Monte Carlo distribution. Barton and David (1966) found this to result in an approximate Poisson distribution but this only applied to that specific case (Williams, 1984). However, with heavy clustering in either the space or time dimension the switching of already close labels tends to influence the variance of the resulting Monte Carlo distribution. Theophilides et al., (2006) presented a different remedy, using a completely random Monte Carlo distribution of points within the spatial and temporal domain of the dataset. The probability of randomness was defined as the proportion of the upper right hand tail of the pair distribution whose number of close space-time, close space and, close time pairs were greater than or equal to the respective actual counts. This test is thus an *unconditional extension* of the Knox test, taking directly into account the distribution of the

marginal totals. While there is widespread belief that disease distribution is rarely, if ever, completely random, this presumption fails to account for the dynamic nature of rapidly spreading phenomena. The spatial and temporal characteristics of disease incidents vary in measurement system and units, and scale. Of these, scale is probably one of the most important defining concepts of Geographic analysis. Yattaw (1999) has presented an organizational framework of geographic movement over time and suggested scale as a defining element of space-time analyses. While the non-randomness assumption may hold for slow appearing disease incidents where the population examined is nearly static (i.e. leukemia), it fails for phenomena that are dynamic and operate on finer temporal and/or spatial scales. In the case of West Nile virus and birds, the temporal scale influences the consideration of bird movement. This in effect influences the cluster analysis in space and time that could otherwise be examined separately. Hence the need for a test of space-time randomness arises. Most of the applications of the Knox test have been performed retrospectively in single areas, but recently Rogerson (2001) has presented a prospective local area version of the test. Theophilides et al., (2003, 2004) implemented the Knox test dynamically over local areas and customized it to model a process, the amplification cycle of West Nile virus in the wild (DYCAST methodology). West Nile virus is a mosquito born agent that under favorable environmental conditions cycles and amplifies between birds and mosquitoes. When the cycle of infection is intense, significant amplification of the virus numbers occurs and it spills-over to humans causing illness in the elderly and the immune-weak (Campbell, 2001). In Theophilides (2006), the ability of the unconditional extension of the Knox test (with a Monte Carlo methodology) to predict human risk for West Nile virus was quantified using a kappa methodology and proved that the method models the amplification and transmission to humans process. Here we use the same methods and data to compare the Monte Carlo unconditional Knox results of Theophilides et al.,(2006) with a chi-squared test in terms of their discriminatory power to "predict" human risk. We further directly evaluate the agreement between the two tests in terms of identifying West Nile risk areas.

2. Materials and Methods

Data were provided by the Chicago Department of Health and included the location of dead birds (mostly crows and blue jays, n = 3837, henceforth called the data points) and their date of reporting, and the location of residence and date of onset of illness of humans (n=215) confirmed with West Nile infection.

2.1 Methods

We followed the DYCAST method of Theophilides et al., (2003, 2006). The DYCAST method partitions space into overlapping local units of ecologically relevant spatial and temporal domains that consist of a 1.5 mile radius and the 21 days prior to the current date. This space-time domain was centered about each 0.5x0.5mile grid cell that was laid across the study area (City of Chicago). A Knox analysis was run using all the data points found within the spatial and temporal domains of each cell centroid for each day between June 30th, 2002 and October 5th, 2002. Two probability values were recorded for each day for each cell; one using a chi-square test and one using a Monte Carlo simulation (Theophilides et al., 2006).

2.2 Knox Test

The Knox test statistic is mathematically expressed as:

$$T(o_{11}) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} s_{ij} t_{ij} \ , \qquad \text{Equation 1}$$

Where: s_{ij} and t_{ij} are 1 if the pair formed by the i^{th} and i^{th} point is close in space (interpoint spatial distance <= s) and time (interpoint temporal distance <= t) respectively, and n is the total number of points found in the geographic and temporal space (Table 1). The critical spatial distance was defined as 0.25 miles (s=0.25m) and the critical temporal distance was 3 days (t=3 days).

Table 1: Knox Space Time Correspondence

Knox matrix	**Space**		
Time		Close	Not Close
	Close	o_{11}	o_{12}
	Not Close	o_{21}	o_{22}

2.3 Chi-Square Test

To assess the significance of the test statistic using a chi-squared test a matrix is formed and the resulting chi-squared value was inverted based on a chi-squared function with one degree of freedom to obtain a probability (Ross, 1999).

2.4 Unconditional Monte Carlo Test

For a number of pairs n found within the Spatial Domain and Temporal Domain, 5000 random sets of n points were generated within the same domain.

For each set, the number of close space-time pairs (st=o_{11}), the number close space pairs (s=o_{11} + o_{21}) and, the number of close time pairs (t=o_{11} + o_{12}) were counted. Based on the distribution of these 5000 random runs the probability level was assessed as:

$$p^u = P[(st \geq ST) \bigcap (s \geq S) \bigcap (t \geq T)]$$

Equation 2

Where S, T, and ST are the actual number of points found close in space, time, and both respectively. This unconditional formulation should be contrasted with that of the traditional Knox test where:

$$p = P[(st \geq ST) \mid (s \geq S) \bigcap (t \geq T)]$$

Equation 3

Hence

$$p^u = p * P[(s \geq S) \bigcap (t \geq T)]$$

Equation 4

2.5 Comparison methodology

In this research we are interested in evaluating the DYCAST results obtained with different statistical tests of the Knox method (chi-square and unconditional Monte Carlo) for assessing (1) the discriminatory power of theses tests to successfully identify areas where human cases occurred prior to their onset of illness and (2) the agreement between them. To accomplish this we use a new implementation of the kappa test which has been adapted for a spatial and temporal domain (Theophilides et al., 2006). The kappa test was introduced by Cohen (1960) as a means of measuring the agreement of classification of data into diagnostic categories between two judges (or raters) after chance agreement was excluded. Later, Congalton and Mead (1983) implemented the kappa over spatial data in Remote Sensing as a means of assessing the accuracy of image classification. The general form of the kappa statistic is:

$$\hat{\kappa} = \frac{N \sum_{\iota=1}^{r} x_{ii} - \sum_{i=1}^{r} (x_{i+} * x_{+i})}{N^2 - \sum_{i=1}^{r} (x_{i+} * x_{+i})},$$

Equation 5

Where: N is the total number of areas considered, and x_{ii}, x_{i+}, x_{+i} are the elements of the following matrix (Table 2): The sum of which amounts to *N*.

Table 2: Kappa matrix

	Rater 1		
Rater 2		Class 1	Class 2
	Class 1	x_{11}	x_{12}
	Class 2	x_{21}	x_{22}

In the spatial and temporal implementation of kappa by Theophilides et al., (2006) a kappa value is calculated for all unique combinations of a range of days prior to onset of illness in humans and a range of temporal windows in order to find the date and duration of maximum non-chance agreement between DYCAST results and human illnesses. The total number of cells for any unique combination of days prior to onset and a window is expressed as:

$$N = \sum_{d=June30th,2002}^{October5th,2002} G * I$$

Equation 6

Where *G* is equal to 1181, the number cells of the Chicago grid and *I* is an indicator function which can be evaluated as follows:

(1) *I = 1* if for dates (*d+p days) to(d+p days +(w-1 days)* there is at least one human case. where: *p* is the number of days prior to onset and *w* the window size.
(2) *I = 0* otherwise.

The calculation of the kappa values and their respective chi-square test was done over a combination of *p=0 to 21* and *w=1 to 19.*

To evaluate direct non-chance agreement between the chi-square test and the unconditional Monte Carlo test, we summed all the cells with agreements (and disagreements) over the whole period and constructed a kappa Table as the one shown before.

3. Results

The results of the non-chance agreement between the human cases and the DYCAST results with the unconditional Monte Carlo method were presented in Theophilides et al., (2006). An example of a results map, showing a specific date of risk calculation with the knox statistic and future identification of risk areas (post-analysis) is shown in map 1.

Table 3: Kappa agreement table

Days prior	Windows 19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
21	-0.01	0.008	0.026	0.044	0.062	0.078	0.096	0.114	0.13	0.141	0.157	0.175	0.192	0.204	0.219	0.255	0.295	0.321	0.358
20	0.037	0.054	0.072	0.091	0.107	0.125	0.143	0.164	0.18	0.197	0.215	0.233	0.251	0.267	0.298	0.333	0.354	0.382	0.381
19	0.082	0.1	0.118	0.135	0.153	0.171	0.192	0.213	0.231	0.25	0.268	0.288	0.307	0.336	0.366	0.387	0.408	0.414	0.41
18	0.124	0.142	0.158	0.176	0.195	0.215	0.236	0.258	0.277	0.295	0.315	0.334	0.361	0.388	0.405	0.423	0.425	0.429	0.419
17	0.166	0.182	0.2	0.218	0.238	0.258	0.28	0.302	0.321	0.34	0.36	0.385	0.409	0.425	0.442	0.446	0.45	0.455	0.456
16	0.205	0.222	0.24	0.26	0.28	0.301	0.323	0.345	0.364	0.384	0.407	0.43	0.445	0.461	0.468	0.475	0.48	0.498	0.503
15	0.237	0.255	0.275	0.295	0.316	0.338	0.359	0.381	0.401	0.423	0.445	0.46	0.474	0.482	0.491	0.499	0.512	0.528	0.526
14	0.27	0.289	0.31	0.33	0.352	0.373	0.395	0.417	0.439	0.46	0.474	0.488	0.496	0.506	0.515	0.529	0.543	0.555	0.564
13	0.302	0.322	0.343	0.364	0.384	0.406	0.428	0.453	0.473	0.486	0.499	0.508	0.517	0.527	0.541	0.554	0.563	0.579	0.577
12	0.334	0.354	0.375	0.395	0.416	0.437	0.461	0.484	0.497	0.51	0.518	0.527	0.536	0.549	0.562	0.57	0.583	0.591	0.584
11	0.363	0.383	0.403	0.423	0.444	0.467	0.488	0.504	0.517	0.524	0.532	0.541	0.552	0.563	0.57	0.58	0.584	0.587	0.566
10	0.39	0.409	0.429	0.449	0.47	0.491	0.506	0.521	0.529	0.536	0.543	0.553	0.561	0.567	0.576	0.578	0.58	0.575	0.558
9	0.413	0.432	0.451	0.472	0.492	0.506	0.52	0.53	0.537	0.543	0.551	0.558	0.562	0.568	0.57	0.571	0.566	0.559	0.535
8	0.435	0.453	0.473	0.491	0.505	0.518	0.527	0.537	0.543	0.549	0.554	0.557	0.561	0.562	0.562	0.557	0.555	0.545	0.526
7	0.453	0.472	0.49	0.503	0.515	0.524	0.532	0.541	0.547	0.55	0.551	0.555	0.554	0.553	0.549	0.546	0.542	0.534	0.503
6	0.471	0.487	0.499	0.511	0.519	0.527	0.535	0.543	0.546	0.546	0.548	0.547	0.545	0.541	0.538	0.534	0.531	0.516	0.49
5	0.484	0.495	0.506	0.513	0.52	0.527	0.535	0.541	0.54	0.541	0.539	0.536	0.531	0.528	0.523	0.52	0.514	0.498	0.468
4	0.49	0.499	0.506	0.513	0.519	0.526	0.531	0.533	0.534	0.531	0.527	0.522	0.517	0.511	0.507	0.5	0.494	0.474	0.441
3	0.493	0.498	0.504	0.51	0.516	0.52	0.521	0.521	0.517	0.513	0.506	0.5	0.493	0.486	0.479	0.471	0.46	0.433	0.393
2	0.491	0.495	0.5	0.505	0.509	0.509	0.507	0.504	0.499	0.492	0.485	0.476	0.469	0.46	0.452	0.441	0.422	0.391	0.359
1	0.488	0.492	0.496	0.499	0.499	0.497	0.492	0.487	0.481	0.473	0.464	0.456	0.447	0.438	0.429	0.412	0.395	0.375	0.365
0	0.483	0.486	0.488	0.488	0.486	0.481	0.475	0.468	0.461	0.452	0.443	0.433	0.424	0.414	0.399	0.383	0.373	0.359	0.332

Table 4: Comparison of Chi-square and Unconditional Monte Carlo

	Unconditional Monte Carlo		
Chi-square		**West Nile risk**	**No West Nile risk**
	West Nile risk	7134	47
	No West Nile risk	10866	97691

In Theophilides et al., (2006) the results were shown as a surface of values for variable size windows and days prior to onset (Figure 1, used with permission). Those results showed that consistent non-chance agreement of 50% occurred for a range of ecologically relevant days prior and windows (Table 1, used with permission) From those results, Theophilides et al., (2006) deduced that peak viremia in the wild birds (corvids) occurs 15-16 days prior to the onset of illness in humans.

The structure of the surface also showed that bird deaths start to decrease after the peak period and consistently decrease up to the date of onset. Theophilides et al., (2006) attributed the human infections to the decrease of birds available for mosquito blood-meals and the subsequent turn of mosquitoes to humans at approximately day 7 prior to onset and proved the hypothesis first posed by Despommier, (2001). A similar surface showing the agreement between the DYCAST chi-square results and the human cases is shown in Figure 2. The surface shows that the kappa values are lower than those of the unconditional Monte Carlo DYCAST results by approximately 0.4 (or a chance agreement difference of 40%). The approximate structure of the surface remains the same but the highest kappa value of 0.1592 occurs for a window of 1 day for 11 days prior to onset. The surface from the chi-square approximation of the Knox statistic retains the structure of gradual increase and decrease of kappa values.

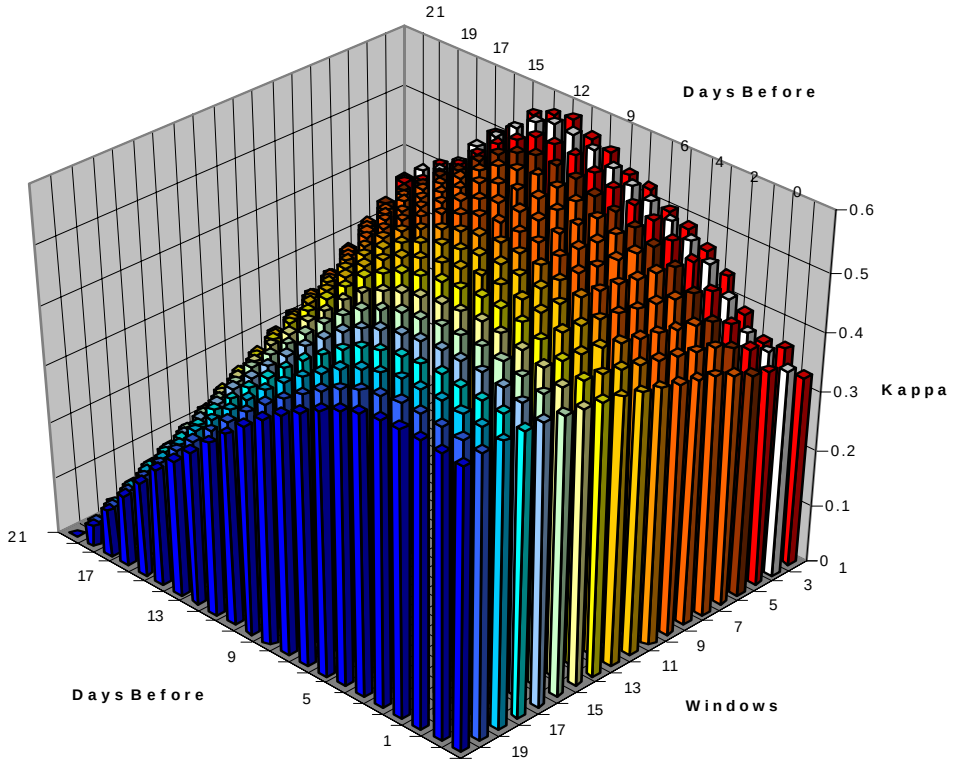


Figure 1: Monte Carlo resulting Kappa value surface

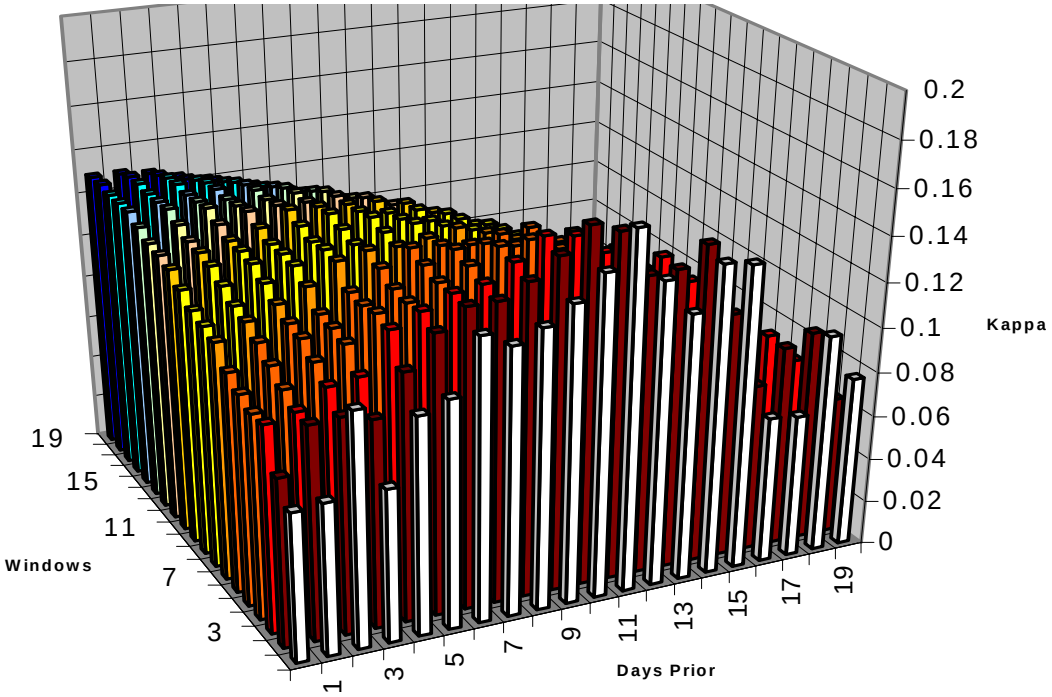
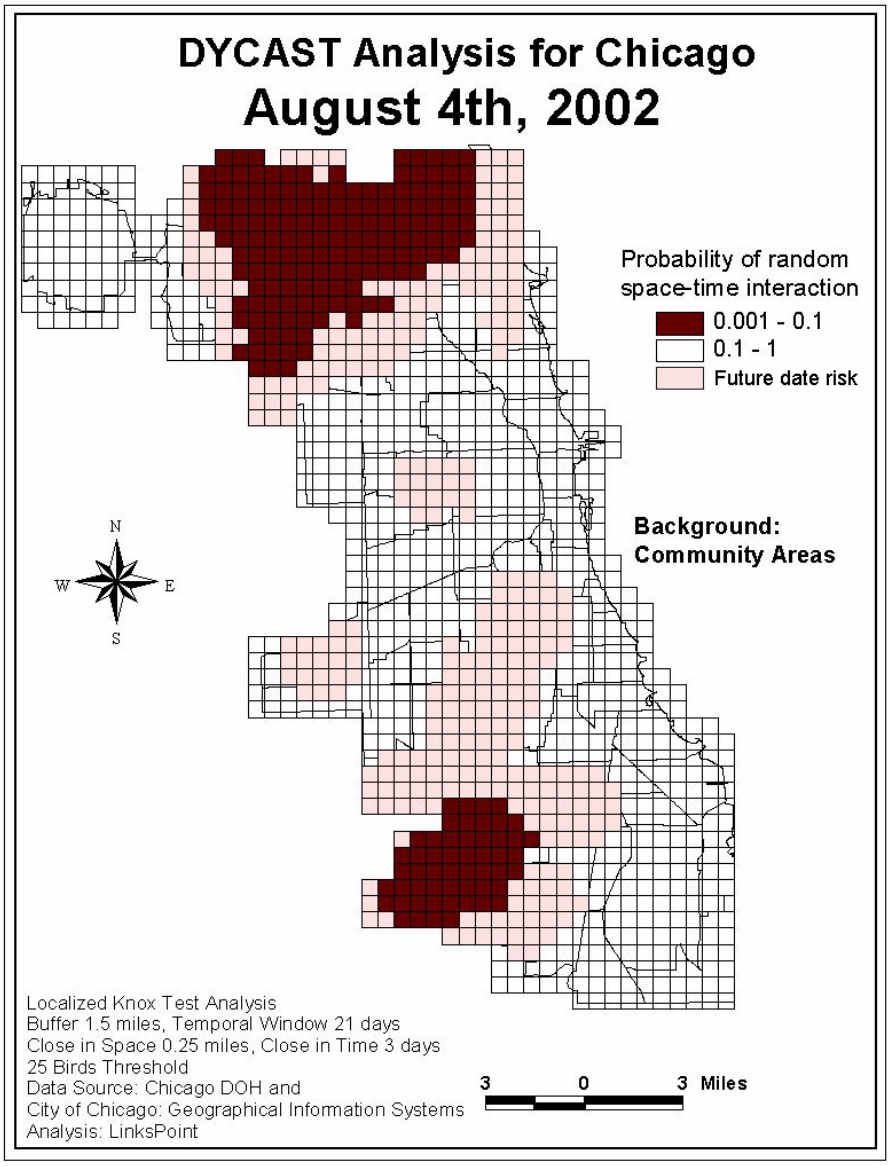


Figure 2: Chi-Sqare resulting in kappa value surface (In different angle from Figure 1.)



Map1: DYCAST analysis for August 4th, 2002: risk areas in dark red, future date risk areas are in pink

As the window decreases from 19 to 1 days, the peak kappa value shifts to higher days prior to onset, verifying the approximate time of highest viremia to be around 15 days prior. However as the windows become smaller and reach two and one days the structure of the surface becomes inconsistent, even though the trend of the surface is retained. This is a visual manifestation of the use of a chi-square test which assumes independence for testing a pair dependent dataset. The same probably occurs for larger window sizes, but its only manifestation then are the lower kappa values. The analysis with large window sizes includes more temporally autocorrelated cells that average out any fluctuations. The direct evaluation of non-chance agreement between the chi-square method and the unconditional Monte Carlo method for testing significance shows a kappa value of 0.52 ($p < 0.001$). This means that only about half the agreement between the two methods is not due to chance (Table 4) and that in general, both tests show the same space and time of bird clustering as significant by a rather high measure of non-chance agreement. Table 4 shows that only 47 cells were shown as at risk with the chi-square method and not with the Monte Carlo method. In contrast, 10886 cells were shown as at risk with the Monte Carlo method and not with the chi-square method. This, combined with the high kappa value with the human cases (Figure 1) is evidence that the Monte Carlo method is a more sensitive measure of evaluating non-random space-time interaction and a more

accurate one. The largest number of cells in agreement between the two techniques was in the not at-risk category (97691, Table 4).

4. Conclusion

We have presented quantitative evidence that the widespread use of the chi-square test for evaluating the significance of space-time interaction of the Knox method (1964) is inappropriate. For smaller granularities of time it renders results with a high degree of variability. Knox (1964) first identified the potential pitfall of this statistic due to pair dependencies and Theophilides et al., (2003) reported this as a result of an indication from the results of a practical application. In the case of using the Knox method with DYCAST, the unconditional Monte Carlo method for evaluating significance, is more sensitive and gives results that more accurately follow West Nile virus activity. This is shown by the rather high kappa measure of non-chance agreement between the human cases and the results (Theophilides et al., 2006). The non-chance agreement of the unconditional Monte Carlo DYCAST results with chi-square DYCAST results is almost 50% although both methods accurately show where activity has not occurred. We caution that more studies may be needed to fully understand the distribution of the Knox statistic both at the theoretical level and for practical applications.

References

Alexander, F. E. 1992, Space-time clustering of childhood acute lymphoblastic leukaemia: indirect evidence for a transmissible agent, *British Journal of Cancer*, 65, 589-592.

Andersson, R., Hugander, A., Thulin, A., Nystrom, P. O. and Olaison, G., 1995, Clusters of Accute Appendicitis: Further Evidence for an Infectious Aetiology. *International Journal of Epidemiology*, 24, 4, 829-833.

Barton, D. E. and David, F. N., 1966, The Random Intersection of Two Graphs. In: David F. N., ed. Research papers in statistics. New York: John Wiley & Sons Inc, 45-59

Heining, R., 2003, *Spatial Data Analysis*: Theory and Practice. Cambridge University Press, Cambridge.

Campbell, G. L, Marfin, A. A., Lanciotti, R. S., and Gubler, D. J., 2002, West Nile virus. *Lancet Infectious Diseases*, 2, 519-529.

Cohen, J., 1960, A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46.

Congalton, R. G., and Mead, R. A., 1983, A quantitative method to test for consistency and correctness in photointerpretation. *Photogrammetric Engineering and Remote Sensing*, 49(1), 69-74.

Despmommier, D. D., 2001, West Nile story. Apple Trees publications, New York

Knox, G., (1963), Detection of low intensity epidemicity: application to cleft lip and palate. *British Journal of Preventive and Social Medicine*, 17, 121-127.

Knox, G., 1964, Epidemiology of childhood Leukemia in Northumberland and Durham. *British Journal of Preventive and Social Medicine*, 18, 17-24.

Knox, G., 1964b, The Detection of Space-Time Interactions. *Applied Statistics* C, 13, 25-30

Lawson, A., B., 2001, *Statistical Methods in Spatial Epidemiology*. John Wiley & Sons, New York.

Mantel, N., 1967, The detection of disease clustering and a generalized regression approach, *Cancer Research*, 27, 209-220.

Rogerson, P. A. 2001, Monitoring point patterns for the development of space-time clusters. *Journal of the Royal Statistical Society* A, 164, 87-96.

Ross, S., 1998, A first Course in Probability, Prentice Hall, New Jersey.

Theophilides, C. N., Ahearn, S. C., Grady, S., and Merlino, M. 2003, Identifying West Nile virus risk areas: The Dynamic Continuous-Area Space-Time system. *American Journal of Epidemiology*, 159, 9, 843-854.

Theophilides, C. N., Ahearn, S. C., Binkowski, E. S., Paul, W. S., and Kibbs, K. 2004, First Evidence of West Nile virus amplification and relationship to human infections. *International Journal for Geographic Information Science* Vol. 20, No. 1, January 2006, 103-115.

Williams, G. W., 1984, Time-Space Clustering of Disease., In Cornell, E. G., *Statistical Methods for Cancer Studies*, Marcel Decker, Inc., New York, 167-227.

Yattaw, N. J., 1999, Conceptualizing Space and Time: A Classification of Geographic Movement. *Cartography and Geographic Information Science*, vol. 26, no. 2: 85-98.