ERDC/CERL SR-09-8

US Army Corps
of Engineers®
Engineer Research and
Development Center

Actionable Cultural Understanding for Support to Tactical Operations (ACUSTO)

Annotated Bibliography for The Effect of Data Quality on Spatial Analysis Results

Luis Galvis and William D. Meyer

June 2009

Construction Engineering
Research Laboratory

ERDC/CERL SR-09-8
June 2009

Actionable Cultural Understanding for Support to Tactical Operations (ACUSTO)

Annotated Bibliography for The Effect of Data Quality on Spatial Analysis Results

Luis Galvis

Arizona State University
Tempe, Arizona

William D. Meyer

Construction Engineering Research Laboratory (CERL)
U.S. Army Engineer Research and Development Center
2902 Newmark Dr.
Champaign, IL 61822-1076

Final Report

Prepared for Headquarters, U.S. Army Corps of Engineers
Washington, DC 20314-1000

Under Work Unit 21 2040

Abstract: The U.S. Army can use spatial data in ways beyond its normal place, if properly acquired and processed. To ensure the analytical quality within spatial data, data to be analyzed must be collected according to proper, data-specific, scientific standards and then properly preprocessed. If this is not done, resulting spatial analysis will suffer to the point that the information is merely anecdotal. Contained in this report is an annotated bibliography of sources which support the research component known as "The Effect of Data Quality on Spatial Analysis Results" for the Actionable Cultural Understanding to Support Tactical Operations (ACUSTO) research project.

Contents

Preface iv

1 Introduction 1

Background 1
Objectives 2
Approach 2
Mode of Technology Transfer 2

2 Assessing Data Quality 3

3 Development of Geographic Methods and Criminology 12

4 Analysis Developments 14

Representation of crimes in space 14
Spatial visualization of crime 14
Assessing patterns of crimes in space 17
Crimes as events: Scale issues 20

5 Conclusion 23

Acronyms and Abbreviations 24

References 25

REPORT DOCUMENTATION PAGE 27

Preface

This study was conducted for the Assistant Secretary of the Army for Acquisition, Logistics, and Technology (ASAALT) under Project 622784AT41, "Military Facilities Engineering Technology," Work Unit 21 2040, "Social-Cultural and Environmental Data Fusion Models."

The work was completed under the direction of the Ecological Process Branch (CN-N) of the Installations Division (CN), Construction Engineering Research Laboratory (CERL). The CERL Project Manager was William D. Meyer. At the time of publication, Alan B. Anderson was Chief, CN-N, and Dr. John T. Bandy was Chief, CN. The associated Technical Director was Dr. William D. Severinghaus. The Deputy Director of CERL was Dr. Kirankumar V. Topudurti and the Director was Dr. Ilker Adiguzel.

CERL is an element of the U.S. Army Engineer Research and Development Center (ERDC), U.S. Army Corps of Engineers. The Commander and Executive Director of ERDC was COL Gary E. Johnston, and the Director of ERDC was Dr. James R. Houston.

1 Introduction

Background

Developing cultural information into cultural knowledge for military operations is predominantly an intelligence activity that takes place within the Military Decision Making Process (MDMP). MDMP includes mission analysis that produces an intelligence assessment, evaluation of courses of action, and re-evaluation of intelligence assessment. Intelligence Preparation of the Battlefield (IPB) is performed before, during, and after the mission analysis phase of the MDMP. Recent Army field manuals and lessons learned documents emphasize the role of Every Soldier as Sensor (ES2) in providing information for IPB. The incorporation of cultural knowledge into IPB is recognized as especially critical for planning and implementing counterinsurgency operations.

In practice, IPB involves collecting data manually or through sensors, coupled with computer analysis by highly trained intelligence analysts. The products produced from these efforts are routinely classified and subsequently unusable by the tactical war fighter operating at the brigade combat team level.

The Actionable Cultural Understanding for Support to Tactical Operations (ACUSTO) project was undertaken to provide a product for enhanced cultural understanding that will be accessible to the tactical war fighter. This is accomplished through a combination of spatial and explicit content analysis of open source news media and other information, to provide cultural understanding in the operational environment (OE) that can be disseminated down to the lowest tactical level. Once the Soldier possesses enhanced cultural knowledge, this will improve his/her ability to recognize and document significant cultural information. Thus, the quality of observations by ES2 regarding cultural factors will improve. This will be accomplished through the ability to discern where and with whom incidents are most likely to occur, based on an understanding of the drivers of spatial pattern linked with socio-cultural knowledge at the neighborhood scale within the brigade combat team area of operation (AO).

ACUSTO research began with a literature review of current cultural data analysis techniques, which are compiled in this report.

Objectives

The objective of this stage of the project was to find and review spatial data analysis techniques that would have potential application to the ACUSTO program effort. To ensure good analytical quality within spatial data analysis, data must be collected in accordance with specific and proper scientific standards and then properly preprocessed.

Approach

The ACUSTO research project is acutely sensitive to the difficulty of acquiring spatial data in hostile environments, where little time would be available to ensure that exacting data collection protocols could be met. To ensure the quality of spatial data to be analyzed in the ACUSTO research project, a study was made of what types of errors could emerge and cause the data quality to be less than acceptable. To accomplish this, a high-quality spatial dataset of homicides in the Watts area of southeast Los Angeles, CA was used. The data was gradually eroded and analyzed across a number of spatial analysis techniques being used in ACUSTO research. The results were documented in the ERDC-CERL Technical Report, *ACUSTO: The Effect of Data Quality on Spatial Analysis Results.*

Mode of Technology Transfer

This report will be made accessible through the World Wide Web (WWW) at URL: http://www.cecer.army.mil

2 Assessing Data Quality

Jacquez, G., and L. Waller. 2000. The effect of uncertain locations on disease cluster statistics. In H. T. Mowrer and R. G. Congalton, eds. *Quantifying Spatial Uncertainty in Natural Resources: Theory and Applications for GIS and Remote Sensing*. CRC Press.

The authors discussed how cluster analysis is very relevant within public health; it allows assessment of whether a given pattern of cases, as points over space, is statistically unusual.

However, models of spatial point processes (that underlie many point-based statistics) rest on the assumption of exact locations, an assumption not easily upheld in real health data. For instance, although there has been a good deal of research on attribute uncertainty, there is little research focusing on location uncertainty.

Such location uncertainty in health data can arise from: (1) use of area centers rather than actual addresses (or, the closest grid node in the case of raster health data), and (2) the nature of the health issue, as the data might have been assigned to a residential address, but the event occurred throughout a journey to other locations. The authors assert that, in reality, locations are an indication of exposure and therefore carry implicit uncertainty, and the effect of that uncertainty on spatial statistics needs to be considered.

To assess such effects, the analysis first considers the most common case of uncertainty—location, due to aggregation at the area centroid. Such uncertainty could comprise four levels: (1) none (where the actual locations are used) (2) county, (3) bi-county, and (4) region. The performance of three cluster statistics is evaluated — by comparing test statistics, distributions, P-values and power — of simulated AIDS data for 500 individuals in Michigan versus those of the actual locations, from which two individuals are chosen as the seeds for the epidemic. The simulation process included three steps: (1) evaluating transmission events, (2) determining transition probabilities, and (3) recording seropositives.

Each of the three space-time cluster statistics used in the analysis has a different measure of spatial proximity. The Mantel's test is based on geographic distance, regressing the space between pairs of cases on their time distance, with a null hypothesis of independence of time and space distances between cases. The reference distribution is generated either through a normal approximation of Mantel or a randomization of the time locations over the case locations.

The Knox test is based on adjacencies (where time and space adjacencies are defined against a critical value) that are likely to be based on the underlying theoretical assumptions of the epidemiological process, with a null hypothesis of independence of space and time adjacencies. The significance is evaluated by randomization or by a chi-square test.

The last test addressed by the authors is the K-NN test, which is based on nearest neighbor relationships, as it considers that fixed critical spatial distance cannot address the variation of population densities. The test considers the intersection between space and time nearest neighbor relationships, through a count of the number of case pairs that are *k* nearest neighbors in time and space. The significance is evaluated through randomization *o* the time locations over the spatial locations.

All distributions were generated through Monte Carlo randomizations, that included 252 AIDS simulations, three clusters, three levels of aggregation, and 249 randomizations.

The results presented evidence of the change in P-values and the decrease of statistical power, due to the employment of centroids instead of locations. The authors argue that insofar as the study tried to resemble the conditions of a real space-time disease model, population densities and small sample sizes, it is worrisome that the results point to a substantial proportion of false negatives. Clusters are not being detected as a result of the effect of location uncertainty on statistical power. The authors point out that centroids are too dispersed and thus violate the Poisson assumptions that underline most point statistics.

Gabrosek, J., and N. Cressie. 2002. The effect on attribute prediction of location uncertainty in spatial data. *Geographical Analysis* 34(3):262–285.

Although studies using kriging methods usually make the assumption that spatial locations are free of errors, the authors address the effect of location uncertainty on statistical inference by incorporating location error in the kriging equation. The analysis obtains location error by adjusting the components of the kriging equation and doing the analysis with the assumption that the spatial process is sampled without location error. The authors use measurement error models to incorporate location errors into the estimation of spatial dependence and spatial prediction.

This paper follows the error-propagation analysis approach developed by Arbia, Griffith, and Haining, specifying a blurring matrix and a distribution for the error, and through Monte Carlo simulations, generates realization of the true map. In doing so, they can later compare the observed with the true values. The authors described the statistical model underlying the location error coming from design model sampling, one of the common sources of location error, where the realized sites might not be equal to the intended sites, which are fixed and known. This would be due to, for instance, inaccuracies in the positioning instruments. Another source of location error mentioned by the authors is the resource model, where the error comes from the sample locations which are defined by the sampled phenomena that are fixed but unknown.

Results from the analysis showed that location errors cause changes in the matrix of covariances between sampled sites, and the vector of covariances between sampled sites and prediction location.

The authors proposed an approach to kriging that adjusts for location error, denoted as KAALE. This approach performs substantially better than kriging without adjustment, when the location error is nontrivial. Once adjustments for location error are in place, optimal kriging coefficients are found through Lagrange multipliers.

Olson, K., S. Grannis, and K. Mandl. 2006. Privacy protection versus cluster detection in spatial epidemiology. *American Journal of Public Health* 96(11):2002–2008.

Given the increasing importance of detecting outbreaks for a given event, cluster detection has gained importance as a tool to describe and analyze crimes in space. At the same time, efforts to maintain and protect people's confidentiality has led to the use of administrative areas as geographic locations instead of the original-latitude longitude coordinates. The authors seek to expand the understanding of the effect that such aggregation approaches have on the corresponding analysis.

Particularly, the authors assessed the effects that the use of blurring techniques on identifiable patient address data might have on the detection of outbreaks. To do so, they added simulated clusters to baseline data on all emergency department visits, between 2000–2005 regarding respiratory illness, using AEGIS-CCT open source software. Patients' addresses were geocoded in ArcGIS and mapped to zip code or census tract, for which centroids were calculated. Separate analyses were performed for census tract and zip code administrative region, but all clusters had 10 points and were placed about 5 km from the center of one hospital.

The analysis focused on assessing the effect of displacing the points from their original location to the zip code or census tract centroid. The clusters vary by the size of their radii (0.5, 1, 2, and 3 km from the zip code centroids and 0.5 and 1 from the census tract centroid), and the number of areas they expand, up to 4. SatScan was used to detect spatial clustering. Additional analyses were performed in SAS® (Statistical Analysis Software), where the dependent variables were percentage of significant spatial clusters (those with p-values smaller than 0.05 according to SatScan), proportion of significant clusters with at least half the simulated points, and original Emergency Room (ER) visits used in the clusters. The independent variables were the regional centroids, with separate analyses for census tract and zip codes, radius size of the clusters, and number of regions the clusters expanded to. The authors employed a generalized estimating equation, to account for the covariance between the two levels of address precision.

Results pointed out that exact coordinates detected more significant clusters than did zip code centroids and census tract centroids. Moreover, when clusters expand to more than one region, exact coordinates yielded

higher proportions of detected clusters. Finally, the authors point to the differences between the two administrative regions, where the use of census tract instead of exact coordinate locations led to a smaller decrease in cluster detection rates, particularly for cases that expanded over fewer census tract areas.

The authors focused on the dispersion across boundaries as a cluster parameter. They acknowledged that other parameters such as population density around the cluster, different shape of clusters, or various distances from the hospital might still need to be addressed.

Armstrong, M., G. Rushton, and D. Zimmerman. 1999. Geographically masking health data to preserve confidentiality. *Stat Med* 18:497–525.

This paper presented an approach to mask geographic data that would allow researchers access to micro-level data, without disclosing the actual locations of individuals. This procedure is proposed as a solution for the problems that originate when individual records are aggregated to protect confidentiality. The authors note that health data is collected with the purpose of detecting disease clusters or evaluating environmental exposure effects on specific diseases. However, when such data is aggregated to large areas for further release, the suitable clustering methods are limited to those for aggregated areas. Another limitation is that when aggregation is performed, the cluster size is limited to the scale of the aggregated area.

Another issue mentioned is the boundaries of the aggregated area in relation size of the cluster; a problem arises when the latter does not encompass the aggregated area boundaries. The authors propose masking geographic records in such a way that some attributes are kept. This could be accomplished by performing rotations in the original data, rescaling the coordinates, or applying small random displacements of data that keep the nearest neighbor distances. The last type of masking would allow users to conduct, for instance, spatial-clustering methods based on nearest neighbor distances, without compromising confidentiality.

Spatial clustering may not be accurately done when the data are aggregated before releasing. Using a Humberside dataset, the authors conducted an experiment generating certain levels of perturbation to the original coordinates. For each level of perturbation, they simulated 500 random patterns. The results showed that for small levels of perturbation, the clustered pattern still is revealed, but, for higher levels it disappears.

However, this type of masking is preferable to affine and aggregation masks, when the data is to be used for analytical purposes.

Zimmerman, D. L., and C. Pavlik. 2008. Quantifying the effects of mask metadata disclosure and multiple releases on the confidentiality of geographically masked health data. *Geographical Analysis* 40(1):52–76.

The authors addressed the feasibility of breaching confidentiality in masked health data through information in the masked metadata or through the availability of different masked versions of the original dataset.

This paper pointed out that masking has been increasingly used and recognized as both effective and necessary to protect individual identity. Masking involves removing or encrypting personal information such as names, addresses, and social security numbers; sampling, swapping or adding simulated data; and for a spatial dataset, modifying x-y coordinates or areal aggregation.

The analysis first involved defining a quantitative measure for the risk of identity disclosure (the area of the confidence region for the true location of the patient). Such a measure would follow the probability model of the dataset mask — the greater the confidence region, the smaller the disclosure risk. The masking does not account for population density, as it is focused on the physical area.

The masking procedure added a perturbation to disguise the true location of each case. This perturbation is derived from a random sampling following either a bivariate normal distribution or a circular uniform distribution, which the authors consider are more commonly used. The analysis then considers the case where dispersion parameters are disclosed or concealed. In addition, the authors include two scenarios in the analysis, one in which case labels are disclosed and one in which they are undisclosed. The paper follows with a quantification of risk disclosure for these different scenarios, through a combination of simulation and statistical theory.

As the authors expected, results indicate that as the confidence regions decrease, the risk of disclosure increases when masked metadata was disclosed — for lower number of masked datasets releases and datasets with smaller number of cases. In contrast, for datasets with a larger number of

cases or higher number of masked releases, the effect of releasing metadata information is trivial.

With regard to the disclosure of case labels, the analysis indicates that the impact on risk disclosure is modest, with some variation based on the distances between cases or the size of the mask's dispersion parameter.

The authors also pointed out that confidentiality for different data releases, using only an aggregation mask, depends on whether the areas are crossed or nested. If the different releases correspond to areas that are nested, then the risk of disclosure depends on the underlying population density in the smaller areal unit mask. If the releases used areas that are crossed, like census tracts and zip codes, risk of disclosure is related to the size of the population in the areas of intersection across the different masked datasets.

However, the authors argue that when data releases include combinations of dataset with areal aggregation masks and perturbations of point locations, considerable risk of disclosure will exist only for those cases near the boundary of the areal unit.

The authors also further discuss the effect of masking on spatial statistical analyses. They addressed this issue by assessing the degree to which the release of multiple-masked versions or masked metadata, using circular normal perturbations, decreases the impact of masking on the quality of maximum likelihood estimation (MLE) of the parameters of a clustered Poisson process. Results point to the conclusion that the release of the masked metadata—specifically the standard deviation used in the perturbation—has an irrelevant effect on the quality of MLE estimation. However, there is a substantial improvement in MLE estimation quality, as the root mean squared error gets smaller with increases in the number of masked releases, particularly from 1 to 2.

Cassa, C., S. Grannis, J. Overhage, and K. Mandl. 2006. A context-sensitive approach to anonymizing spatial surveillance data: Impact on outbreak detection. *Journal of the American Medical Informatics Association* 13(2):160–165.

The authors present a methodology that deals with the issue of breaching patient privacy and anonymity in location-related information. Baseline control data on ER visits for respiratory illness were combined with artifi-

cial clusters that vary in terms of magnitude, shape and location. These locations were then made anonymous by transposing them through use of an algorithm that skews the location and takes into account the local underlying population density. The algorithm randomizes the magnitude of the address skew for each observation by using a random seed parameter. Such a parameter inversely varies with the underlying population density, to later choose an x and y offset value following a Gaussian probability distribution.

The authors conclude that a population-density-based Gaussian spatial skew algorithm effectively guards the identity of individual subjects in a dataset, while marginally affecting the spatial cluster detection sensitivity of SaTScan. This addresses subject confidentiality while preserving the precision and therefore the sensitivity to cluster detection, information pivotal in outbreak detection by health agencies. Such information is lost when the alternative de-identification by aggregating points is used,.

Cressie, N., and J. Kornak. 2003. Spatial statistics in the presence of location error with an application to remote sensing of the environment. *Statistical Science* 18(4):436–456.

The authors address the issue of spatial inference in the presence of measurement errors at spatial locations. The paper describe the effect of uncertainty in locational information in geostatistics. Presence of location errors affects the covariates at a location, the spatial lag between locations, and consequently, the variogram (spatial covariance function), potentially influencing trend parameter estimation.

The authors argue that, despite some efforts to incorporate location error in geostatistical analysis, some issues still are not addressed. Two such issues are: (1) deriving not only the covariance between observations affected by location error, but also deriving the variances of the individual observations; and (2) including a component of variation for the general trend term in the kriging equations.

The authors differentiate the effects of uncertainty due to location errors coming from the coordinate-positioning model (or design model), and those due to the feature-positioning model (or resource model). In uncertainty of the first type, coordinates for locations are predetermined and errors arise due to instrument imprecisions, with the location-error distribution usually centered around the intended locations. Conversely, in lo-

cation uncertainties of the second type, locations are unknown until the desired features/resources are identified and the measurements arise between the actual location of the feature and the one recorded, with the location-error distribution centered around the unknown locations.

The paper seeks to extend previous work, particularly that of Gabrosek and Cressie (2002), on a coordinate-positioning, location error, geostatistical model. Such a model assumes that location error, measurement error, and spatial-process error are mutually independent, and also that the random process (from which the dataset is a single partial realization), has a linear trend in spatial covariates.

The paper then presents the framework for estimating the adjusted mean, covariance, and variance, as well as the Monte Carlo integration algorithm, to evaluate these three moments adjusted for location errors. They illustrate this framework through an artificially generated process. The location error is introduced through simulation, where the perturbation is obtained from an independent sampling, with a density function derived from a uniform distribution. Also, a simulation experiment and an analysis of total column ozone (obtained through remote sensing) are used to look at the effects of acknowledging location uncertainty.

To obtain the adjustments for location-error, the analysis involves integrals with respect to the location-error density, and in the case of a combination of discrete and continuous location-error distributions, Riemann-Stieltjes integrals.

The authors conclude that there are important gains in efficiency when location error is adjusted in the kriging models. Moreover, the authors argue that, given the importance for the linear trend parameter in their analysis, greater efficiency gains can be expected for non-linear complex trends.

3 Development of Geographic Methods and Criminology

Craglia, M., R. Haining, and P. Wiles. 2000. A comparative evaluation of approaches to urban crime pattern analysis. *Urban Studies* 37(4):711–729.

Geographic information system (GIS) technology is being used not only by police authorities for visualization purposes and neighborhood mapping, but also at the research and planning levels. Given that public agencies need to share information, the efficient use of common geographical data held by different agencies is crucial. Thus, GIS is an ideal way to analyze the distribution of problems in local areas so that initiatives can be targeted effectively. Furthermore, use of crime mapping at the local level may create greater involvement by local communities in crime prevention.

However, problems arise when geocoding crime data, not because of a lack of accurate description of locations, but rather, the difficulty of a location being recognized by geocoding software. The authors mention an array of solutions such as standard gazetteers, automatic geocoding to the full postal address, developing geocoding software to recognize colloquial descriptions, or creating standardized reporting frameworks to map directly to digital maps. Overall, the authors point out the importance of developing a local spatial data infrastructure with the participation of all the key stakeholders in the city.

Chen, H., J. Schroeder, R. Hauck, and L. Ridgeway. 2003. COPLINK Connect: Information and knowledge management for law enforcement. *Decision Support Systems* 34:271–285.

Access, not lack of information, is a critical issue in the timely retrieval of data by law enforcement offices. Data is usually not synchronized among information systems or agencies. The authors pointed to progress being made toward information systems that are easily accessed, not only internally by a given agency, but also those allowing efficient interagency exchange. This is evident in initiatives like the National Incident Based Reporting System (NIBRS) that establishes Federal standards. However, many of these efforts address mostly aspects of information management and data sharing, while tools for analysis are less developed. The authors

present the COPLINK Connect initiative, which seeks to allow information sharing among agencies, under a unified user interface with a relational database. The main advantage of this system is the data integration within and between agencies, and the short time frame for responding to queries. However, it is mainly a tool to access data across agencies, and lacks geo-mapping components or advanced queries with association or linkages capabilities.

Brown, D., and S.Hagen. 2003. Data association methods with applications to law enforcement. Decision Support Systems, 34:369–378.

The authors considered three categories for current data-association tools that support crime analysis: (1) expert systems, (2) investigative support systems, and (3) non-automated crime analysis methods. The authors proposed a method that addresses the data association needed for crime analysis. Their proposal has the following advantages over other data-association tools: (1) allows criminal suspects as well as incidents to be considered; (2) ranks, as well as categories, can be obtained for either criminal suspects or incidents; and (3) rules for the association of data are accessible by the user.

4 Analysis Developments

Representation of crimes in space

Black, W., and I. Thomas. 1998. Accidents on Belgium's motorways: A network autocorrelation analysis. *Journal of Transport Geography* 6(1):23–31.

This paper presented an extension of spatial autocorrelation for flows on a network. The paper's focus was not on a single segment with a high accident count (which can be addressed by a local policy), but rather on lower accident counts that occur along different, possibly contiguous, segments. Therefore, the authors paid attention to segment connectivity rather than spatial proximity. They presented an empirical exercise for the Belgium motorway system, where highway segments, or "blackspots," are identified where there have been an unusually high number of accidents. The authors tested the null hypothesis — that the distribution of accidents or accident rates on segments is not autocorrelated. Results indicated that accident rates were dependent on contiguous segments. Later, two segments were identified as the major source of this positive network autocorrelation.

Spatial visualization of crime

Corcoran, J., G. Higgs, C. Brunsdon, A. Ware, and P. Norman. 2007. The use of spatial analytical techniques to explore patterns of fire incidence: A South Wales case study. *Computers, Environment and Urban Systems* 31:623–647.

This paper sought to demonstrate the importance of spatial analytical approaches to the understanding of fire incidents, an area that has received far less attention in the United Kingdom than crime analysis has received. The paper paid special attention to the relationship between certain types of fire incident callouts and the underlying socio-economic variables of a given location. Townsend scores—an index based on the country's 2001 census—was used as a measure of social deprivation. Other, individual socio-economic and demographic census variables were also used.

In addition, the paper examined the spatial association between different types of callouts, by detecting hotspots through kernel density functions,

very common in crime studies. Data for fire callouts (over a 4-year period at the census block level in Wales) were used. Data included callouts for property, vehicle, secondary and false alarm fires. Incident locations then were aggregated (by type of incident) at the ward level, to obtain a rate of incidence per 1,000 population.

The authors cited previous studies which support the association between fire incidents and spatial variations of another variable. For instance, educational attainment and population density are related to fire risk factors, and deprivation indexes are related both to residential fires and to malicious fire incidents.

Similarly, the authors mentioned studies showing how the strength of the correlation between location and fire rates was mediated by the unit of analysis specified. Their analysis included an initial spatial visualization of the incident rates at the ward level (per 1,000 population). Also, risk surface maps were obtained from kernel density estimations (using a bandwidth parameter of 2 km). Authors emphasized the kernel is based on the physical area, since it is difficult to establish the population at risk in the case of the vehicle and malicious callouts, in contrast with property fire callouts. Principal component analysis was used, to avoid correlation between explanatory variables, and resulted in 32 variables, grouped in seven factors. The paper then followed up with two models (a Poisson and a negative binomial model) for each type of incident callout.

Results from the visualization exercises indicated less deprived wards were associated with lower risk across all incident types. Similarly, as expected due to higher concentrations of population, more incidents are observed toward the southern wards.

Furthermore, the authors argued that, according to the regression results, the Poisson model tended to underestimate the variance of the response variable. They believed this was due to their assumption of non-clustering, leading to an underestimation of standard errors for the coefficient estimates, and inflated z-scores. Thus, the authors proceeded to discuss only the results from the negative binomial model. These results showed that wards with lower levels of education and a higher proportion of white residents were more likely to have a property fire callout, while larger households were more likely to have a vehicle fire callout. False

alarm fire callouts seemed to be associated with wards having a lower proportion of households with children and car owners.

Brunsdon, C., J. Corcoran, , and G. Higgs. 2007. Visualising space and time in crime patterns: A comparison of methods. *Computers, Environment and Urban Systems* 31:52–75.

The authors discussed the contribution of geo-visualization approaches in assessing space-time patterns of crime data, particularly map animation, comap, and isosurface techniques. Efforts to understand the structure of crimes across space have resulted in an array of methods and tools of spatial analysis which are capable of working with disaggregated data. Currently, additional efforts are focused on understanding how these spatial patterns of crime behave over time. The authors noted the advantages of animation techniques to discriminate the spatial distribution of a given phenomena at different time periods, such as crime event peaks at a given moment that can be linked to specific conditions occurring during such period. The paper assessed the three visualization techniques by using data (from 1999 to 2001) on police callouts to disturbances in a 10-mile urban center. All three techniques employ kernel density estimation to produce a risk surface map from the event data.

A Visual Basic®-based software was used for the animation technique. To allow user interaction, this included using standard playback controls, recording options, and map controls to select a given area. After choosing the time scale, risk surface maps for each time period were obtained through ArcView Animal Movement extension, which were then embedded in the animation window. The authors pointed out that the animation identified key risk areas and times of day that had a higher number of disturbances. The paper highlighted the importance of expert, insider knowledge (such as police officers or crime specialists) in maximizing the usefulness of the animation routine.

The second technique, the comap (an extension of coplot), is based on conditional plots of two variables as their relationship changes given the values of a third variable. The variable for this particular analysis was time, which allowed visualization of changes not only between hours, but also between days. Given the number of cases in the study, the authors chose kernel density plots.

The third alternative, the isosurface technique, also allowed daily or longer time scales. The technique consisted of obtaining a probability density function for the likelihood of a disturbance event at a given location and time, through use of a three-dimensional density kernel. This was then visualized through an isosurface map. However, the authors considered that the visualization would become difficult, as isosurface of a given density might also enclose others, preventing them from being observed.

Finally, the authors used linked plots, depicting incident times and locations, with different windows that can be highlighted simultaneously. Again, the authors noted the potential complexity in finding variations of space and time.

Assessing patterns of crimes in space

Cohen, J., and G. Tita. 1999. Diffusion in homicide: Exploring a general method for detecting spatial diffusion processes. *Journal of Quantitative Criminology* 15(4):451–493.

This paper extended a cross-sectional view of the spatial distribution of events. The authors considered spatial dependencies over time and the detection of spatial diffusion processes. Specifically, they focused on empirically exploring the notion that the growth in homicide rates has epidemic-like characteristics, particularly for African-American males. This required a shift to a time-space analysis, because the term “epidemic” implies sudden changes over a period of time, beginning with an initial, rapid growth that is followed by slower declines. The authors utilized and extended Local Indicator of Spatial Association (LISA) methods for spatial autocorrelation to assess the spatial diffusion of homicides in Pittsburgh for the period of 1991 to 1995.

The paper delineated a framework of diffusion that was applicable to homicides and identified contagion as the standard mechanism underlying the diffusion process. They further distinguished between two types of contagion diffusion: (1) relocation diffusion, where the spreading of the process follows the displacement of crimes across different locations while the agents may be the same; and (2) expansion diffusion, where crimes spread from a center that continuously experiences high crime rates. A second mechanism for diffusion process, hierarchical diffusion, is also discussed by the authors. In hierarchical diffusion, diffusion is achieved by a sequence of places, categories, or groups; it does not require direct con-

tact, as imitation or innovation might spur the diffusion in different locations.

The authors evaluated whether sharp increases in citywide homicide totals are accompanied by any systematic spatial diffusion (or spread) of homicide across different neighborhoods in a city. They did this by evaluating successive years of outlier clusters for homicides. This identified the full array of possible combinations of local-neighbor pairs over successive time periods. The various combinations of pairs in successive time periods are each compatible with a different type of diffusion (contagious diffusion — expansion or relocation increase, and hierarchical diffusion —global or isolated increase). The authors assessed large changes by using a Euclidian distance between successive local-neighbor pairs. Changes over time considered to be significant are those that point to a move of at least two standard units in the value of a local-neighbor, LISA pair.

They also evaluated whether transitions are more likely than expected, based on the prevalence of other transitions to the same local-neighbor outcome. This involved reviewing the excess of significant transitions between Low-High youth gang homicide rates at time *t* to High-Low or High-High rates at time *t*+1, relative to the other nonstationary, significant transitions. A *t*-test of the difference between proportions evaluated whether the diffusion transition occurred more often than might be expected, based on the prevalence of other transitions to the same outcome.

Results provided support for the contagious diffusion of increasing homicide rates across neighboring tracts, but only during the year of peak growth in total homicides, when high local rates of youth-gang homicides are followed by significant increases in neighboring youth-nongang rates.

Canter, D., T. Coffey, M. Huntley, and C. Missen. 2004. Predicting serial killers' home base using a decision support system. *Journal of Quantitative Criminology* 4(16):457–478.

The authors sought to expand an understanding of a criminal geographical behavior. Their approach was to evaluate Dragnet, a tool for locating the geographical base of serial offenders. This tool is based on the Circle Hypothesis, which posits that serial offenders are very likely to live within an area that has boundaries defined by their offenses. Thus, geographical behavior can be modeled, once a series of crimes have been adequately linked to a particular offender. The authors mentioned that in the litera-

ture, there is a distinction made between an offender whose crimes are anchored to a given location and an offender whose crimes are anchored to particular type of victim or opportunity. The paper focused on offender-location connection, specifically by modeling the location of offenses in which the base location is the residence of the offender.

Geographic models for this type of criminal investigation include the use of distance decay functions that are used to find the centroid, or center of gravity, of the offender. This particular model is based on research that finding that as distance from the offenders' home increases, the probability of an offense decreases. For the purpose of evaluating the cost-effectiveness of each of the 19 functions used to describe the distance decay, the authors opted to use a rectangle to define the search area, instead of the standard circle with a diameter defined by the distance between the two offenses which are farthest from each other. Also, the search area was increased by 20 percent, to allow the home base to be located outside the rectangle defined by the offenders' crimes.

The authors described the effectiveness of the search according to the weighting assigned to each point within the search area. Such weighting would indicate the likelihood of residence, with equal weighting for tied values. These weightings are used as an index (base index) according to a given decay function. After this, the offender's residence is sought within an array of locations, starting with the highest value of the base index. The cost, a value between 0 and 1, is determined by the proportion of all possible locations searched, before the location of the offender's home base is identified. Additionally, they introduced buffer zones, by inserting step zones of varying sizes in front of the exponential functions, which led to 285 as the final number of functions.

Finally, to allow comparisons among offenders, the authors normalized by the distance between offenses. Two approaches were used for comparison purposes: (1) the mean interpoint distance between all offenses (MID), and (2) the QRange, an index developed by the authors and based on the notion developed in previous studies that the arterial pathways followed by offender might be key to determining the locations of his offenses. The authors argued that, while the MID gives equal weight to all distances, the QRange is the mean perpendicular distance of all offenses and points to the regressional axis, calculated as the linear regression of the crime scene coordinates within an offense distribution.

Crimes as events: Scale issues

Shi, X. 2007. Evaluating the uncertainty caused by post office box addresses in environmental health studies: A restricted Monte Carlo approach. *International Journal of Geographical Information Science* 21(3):325–340.

The author discussed the problem of uncertainty that arises when post office (PO) box addresses are assigned to the closest postal code polygon centroid, and the effect of this uncertainty on cluster analysis. The problem is important because in environmental and health data, the imprecision of patient address is ubiquitous. This is especially true for locations in rural areas, where addresses are typically provided as PO boxes.

The authors pointed to the importance of the analysis of spatial clusters in environmental and health phenomena, particularly the identification of hot spots, which can be the focus of posterior analysis. The paper followed the Jacquez and Jacquez methodology for location models, and Monte Carlo models for randomization.

This paper described the most common methods to deal with hybrid datasets, where an individual point-level address is combined with points from polygon-level addresses (such as centroids). One alternative is to aggregate the point-level observations to the polygon level. This is done by assigning them to the closest centroid or proceeding with a polygon cluster analysis. In the first case, the paper cited previous studies showing that p-values differ substantially from those of the point-level locations. In the second case, the problem of ecological fallacy is present. The other alternative for dealing with hybrid datasets is to disregard the imprecise addresses. However, it is likely that this would lead to bias in the analysis, because PO boxes are not randomly placed; they occur far more commonly in rural areas. A third alternative mentioned in the paper is assigning PO boxes to random polygon-level addresses within the appropriate polygon —thus, introducing uncertainty. The authors built on this last alternative by using a restricted Monte Carlo simulation to estimate the uncertainty introduced by the randomization.

The authors opted for a field cluster method, specifically, a variant of the Geographical Analysis Machine (GAM), in which the clusters are assumed to occur anywhere over a continuous field. Such analysis first involves calculating a density value for each x-y intersection in a grid superimposed on

the study area, based on the location of the actual cases. Then, the significance of the clusters is evaluated through Monte Carlo simulation, by obtaining density estimations for a number of location randomizations and comparing them with the densities for the actual case locations. Here, p-values can be derived from the ranking obtained from the values of actual locations among all locations. The randomization follows the spatial distribution of expected counts based on demographic information on the neighboring area, and the normal rates of disease as defined in epidemiology methods.

The authors extended this global Monte Carlo simulation by performing a restricted Monte Carlo simulation, in which only the polygon-level addresses were randomized. This randomized simulation was only made to the corresponding polygon (in this case, the zip-code polygons), and controlled by the expected counts. As the density values for each actual location vary, the variance of the p-values indicated the uncertainty in the *c* cluster analysis.

In the study referenced by the authors, the empirical exercise for lung-cancer case data for Grafton County, New Hampshire, included 381 cases, with 44 percent matched to point-level locations and the remaining to zip-code level. The authors used Microsoft Visual C++® software to run the Monte Carlo simulations and statistical estimations. Results showed that when there is a significant number of polygon-level addresses in the dataset, an important proportion of the clusters is identified as carrying a low degree of certainty. Given the effect of bandwidth in the results, the authors presented a map with an overlay of the high-certainty areas for three different bandwidth values, identifying the clusters that are consistent across different criteria. The authors pointed out that the relationship between the overall uncertainty and the percent of PO box address is not linear, because it is influenced by the spatial distribution of PO boxes.

Craglia, M., R. Haining, and P. Wiles. 2000. A comparative evaluation of approaches to urban crime pattern analysis. *Urban Studies* 37(4):711–729.

The focus of this paper was the comparison of different methodologies for crime analysis at a citywide level. Specifically, the authors analyzed spatial clustering of domestic burglaries in the center of Sheffield, UK. It is in this area of about 100 km, where 85 percent of this type of burglary occurs within the city. The authors evaluated the Spatial and Temporal Analysis

of Crime package (STAC) and SAGE (for Windows® software environment) cluster detection methods. With regards to the STAC tool, results are dependent on the search radius chosen, and the ranking of the cluster is affected by the ratio between the number of events and the area of the ellipse. Moreover, STAC assumes a uniform distribution of population within clusters. On the other hand, SAGE (which tightly couples Arc/Info with a statistical suite) is directed to assess relative risk, particularly in targeting epidemiology issues. This includes standardization by a set of variables such as age, gender, and urbanization, and adjustments for small population. Clusters were detected by Getis-Ord statistics. In contrast to the STAC results, SAGE was able to identify two smaller clusters that the STAC could only identify when a smaller radius was employed. Although STAC is useful in assessing the clustering of raw count crime data, it does not allow for evaluating the relative risk nor for identifying areas that have a higher or lower crime rate than what it is expected according to a set of factors. Moreover, STAC is only pertinent to identifying a cluster when there is global clustering, as shown by the nearest-neighbor routine (because it consists of computing an average measure for the area). On the other hand, the Getis-Ord routine of the SAGE tool can test for local clustering, but it rests on an assumption of randomness (or, not overall clustering). Finally, the authors highlighted the advantage of the Getis-Ord over the STAC technique with regard to the scale used. The Gertis-Ord technique considers distance bands and, thus, allows testing for clustering at a range of distance scales; the STAC technique tests for clustering at only one scale and distance, to the nearest neighbor.

5 Conclusion

A broad literature search of cultural information analysis was completed, and the resulting annotated bibliography reviewed a number of analysis techniques which have been applied to cultural information. This report provides a resource for ACUSTO researchers, furthering their work of relating cultural information to the tactical fighter level, for better preparation of the battlefield and enhanced mission success.

Acronyms and Abbreviations

Term	Spellout
ACUSTO	Actionable Cultural Understanding for Support to Tactical Operations
AIDS	acquired immune deficiency syndrome
ALT	Acquisition, Logistics, and Technology
AO	area of operation
ASA	Assistant Secretary of the Army
CERL	Construction Engineering Research Laboratory
ER	Emergency Room
ERDC	Engineer Research and Development Center
ES2	Every Soldier as Sensor
GAM	Geographical Analysis Machine
GIS	geographic information system
IPB	Intelligence Preparation of the Battlefield
KAALE	Kriging Approach that Adjusts for Location Error
LISA	Local Indicator of Spatial Association
MDMP	military decision making process
MID	mean interpoint distance
MLE	maximum likelihood estimation
NIBRS	National Incident Based Reporting System (NIBRS)
OE	operational environment
PO	post office
SAS	Statistical Analysis Software
SAGE	Software that couples Arc/Info with a statistical suite
STAC	Spatial and Temporal Analysis of Crime
TR	Technical Report
UK	United Kingdom
URL	Universal Resource Locator
WWW	World Wide Web

References

Armstrong, M., G. Rushton, and D. Zimmerman. 1999. Geographically masking health data to preserve confidentiality. *Stat Med* 18:497–525.

Black, W., and I. Thomas. 1998. Accidents on Belgium's motorways: A network autocorrelation analysis. *Journal of Transport Geography* 6(1):23–31.

Brown, D., and S. Hagen. 2003. Data association methods with applications to law enforcement. *Decision Support Systems* 34:369–378.

Brunsdon, C., J. Corcoran, and G. Higgs. 2007. Visualising space and time in crime patterns: A comparison of methods. *Computers, Environment and Urban Systems* 31:52–75.

Canter, D., T. Coffey, M. Huntley, and C. Missen. 2004. Predicting serial killers' home base using a decision support system. *Journal of Quantitative Criminology* 4(16):457–478.

Cassa, C., S. Grannis, J. Overhage, and K. Mandl. 2006. A context-sensitive approach to anonymizing spatial surveillance data: Impact on outbreak detection. *Journal of the American Medical Informatics Association* 13(2):160–165.

Chen, H., J. Schroeder, R. Hauck, and L. Ridgeway. 2003. COPLINK Connect: Information and knowledge management for law enforcement. *Decision Support Systems* 34:271–285.

Cohen, J., and G. Tita. 1999. Diffusion in homicide: Exploring a general method for detecting spatial diffusion processes. *Journal of Quantitative Criminology* 15(4):451–493.

Corcoran, J., G. Higgs, C. Brunsdon, A. Ware, and P. Norman. 2007. The use of spatial analytical techniques to explore patterns of fire incidence: A South Wales case study. *Computers, Environment and Urban Systems* 31:623–647.

Craglia, M., R. Haining, and P. Wiles. 2000. A comparative evaluation of approaches to urban crime pattern analysis. *Urban Studies* 37(4):711–729.

Cressie, N., and J. Kornak. 2003. Spatial statistics in the presence of location error with an application to remote sensing of the environment. *Statistical Science* 18(4):436–456.

Gabrosek, J., and N. Cressie. 2002. The effect on attribute prediction of location uncertainty in spatial data. *Geographical Analysis* 34(3):262–285.

Jacquez, G., and L. Waller. 2000. The effect of uncertain locations on disease cluster statistics. In *Quantifying Spatial Uncertainty in Natural Resources: Theory and Applications for GIS and Remote Sensing*, eds. H. T. Mowrer and R. G. Congalton, 53-64. Boca Raton, FL: CRC Press, of Taylor & Francis Group.

Olson, K., S. Grannis, and K. Mandl. 2006. Privacy protection versus cluster detection in spatial epidemiology. *American Journal of Public Health* 96(11):2002–2008.

Shi, X. 2007. Evaluating the uncertainty caused by post office box addresses in environmental health studies: A restricted Monte Carlo approach. *International Journal of Geographical Information Science* 21(3):325–340.

Zimmerman, D. L., and C. Pavlik. 2008. Quantifying the effects of mask metadata disclosure and multiple releases on the confidentiality of geographically masked health data. *Geographical Analysis* 40(1):52–76.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
June 2009	Final	

4. TITLE AND SUBTITLE
Actionable Cultural Understanding for Support to Tactical Operations (ACUSTO): Annotated Bibliography for The Effect of Data Quality on Spatial Analysis Results

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT

6. AUTHOR(S)
Luis Galvis and William D. Meyer

5d. PROJECT NUMBER
622784AT41

5e. TASK NUMBER

5f. WORK UNIT NUMBER
21 2040

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army Engineer Research and Development Center (ERDC)
Construction Engineering Research Laboratory (CERL)
PO Box 9005,
Champaign, IL 61826-9005

8. PERFORMING ORGANIZATION REPORT NUMBER
ERDC/CERL SR-09-8

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Assistant Secretary of the Army for Acquisition, Logistics, and Technology (ASAALT)
103 Army Pentagon
Washington, DC 20310-0103

10. SPONSOR/MONITOR'S ACRONYM(S)
ASAALT

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The U.S. Army can use spatial data in ways beyond its normal place, if properly acquired and processed. To ensure the analytical quality within spatial data, data to be analyzed must be collected according to proper, data-specific, scientific standards and then properly preprocessed. If this is not done, resulting spatial analysis will suffer to the point that the information is merely anecdotal. Contained in this report is an annotated bibliography of sources which support the research component known as "The Effect of Data Quality on Spatial Analysis Results" for the Actionable Cultural Understanding to Support Tactical Operations (ACUSTO) research project.

15. SUBJECT TERMS
bibliography, ACUSTO, data management, spatial data analysis

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER (include area code)
Unclassified	Unclassified	Unclassified		32	