

Leukemia clusters in upstate New York: how adding covariates changes the story

Christina Ahrens¹, Naomi Altman², George Casella², Malaika Eaton³, J. T. Gene Hwang⁴,
John Staudenmayer^{1*†}, Catalina Stefanescu¹

¹*School of Operations Research, Cornell University, Ithaca, NY 14853, U.S.A.*

²*Department of Biometrics, Cornell University, Ithaca, NY 14853, U.S.A.*

³*Law School, Cornell University, Ithaca, NY 14853, U.S.A.*

⁴*Department of Mathematics, Cornell University, Ithaca, NY 14853, U.S.A.*

SUMMARY

The 1978–1982 New York State TCE / Leukemia dataset is often used to test new cluster detection methodologies. We augment that dataset with demographic covariates from the 1980 census and find evidence that the relation between several of the TCE wastesites and elevated leukemia rates is probably confounded by the population's age and employment characteristics. This demonstrates a problem that is often mentioned, but seldom touched in detail—clustering can be related to covariates not directly related to the risks of interest. Copyright © 2001 John Wiley & Sons, Ltd.

KEY WORDS: clustering; covariates; confounding; environmental epidemiology

1. INTRODUCTION

Searching for clusters of disease around putative sources has long fascinated epidemiologists and biostatisticians. While disease clusters may be associated with environmental hazards, use of clusters to infer causality is problematic. In this article we re-examine a dataset of Waller *et al.* (1992) (and Waller's thesis Waller, 1992 using data from Iwano, 1989) which was used to examine the relationship between trichloroethylene (TCE) waste sites and leukemia in upstate New York (1978–1982). Our analyses of these data illustrate one pitfall, the presence of confounding variables. The original and subsequent authors recognized the possibility that their results were confounded, but they did not investigate further.

The paper consists of three sections. In the remainder of this section we summarize the analyses of Waller *et al.* (1992) and subsequent re-analyses. In Section 2 we augment Waller's data with additional covariates, fit a generalized linear model and discuss the results. The final section summarizes our results and their implications, describes the current status of the sites identified as possible disease foci, and discusses the relation between environmental statistics and public policy.

*Correspondence to: J. Staudenmayer, Department of Biostatistics, Harvard School of Public Health, 655 Huntingdon Avenue, Boston, MA 02115, U.S.A.

†jstauden@hsph.harvard.edu

Contract/grant sponsor: NIEHS

Contract/grant number: ES07261

Received 15 December 1999

Accepted 11 March 2001

1.1. Waller *et al.* (1992)'s data

This subsection describes the data and the tests used in Waller *et al.* (1992). Note that the data and several analyses also appear in the first chapter of the book *Case Studies in Biometry*, Lange *et al.* (1994).

TCE is an industrial solvent, suspected of contributing to leukemia incidence in exposed individuals. Direct manufacturing contact and ground-water infiltration are common exposure vectors. Although the verdict is still out on whether or not TCE is carcinogenic (Kimbrough *et al.*, 1985), TCE can be seen as a proxy for dangerous industrial contamination since it is often stored with other volatile organic compounds (Waller and McMaster, 1997).

The exposure data include the locations of 11 TCE waste sites compiled by the New York Department of Environmental Conservation. These are listed in Table 1 and on the map in Figure 1.

The outcome data include the number of incident leukemia cases, the population and the location of the 790 census regions in the counties of interest. The census regions include block groups for seven of the counties and census tracts for Broome County due to geocoding difficulties (Waller *et al.*, 1992). The database of cases and addresses was built by the New York State Department of Health and contains all reported cases in the area of interest from 1978 to 1982 (Iwano, 1989). Since the address of some cases could only be narrowed down to the county or census tract level, some cases were fractionally allocated to several block groups by the population in those block groups (Waller, 1992). The source for the population data was the 1980 U.S. census.

1.2. Application of general and focused tests

Waller *et al.* (1992) apply several versions of two types of tests of spatial randomness to their data: *general tests* and *focused tests* (Besag and Newell, 1991). Both types of tests address the same null hypothesis (H_0): every person is equally likely to contract the disease independently of other cases and of the location of his or her residence.

The underlying probabilistic model considers a study region divided into I subregions with population size n_i in each subregion $i = 1, \dots, I$. For every $i = 1, \dots, I$, let C_i be a random variable representing the number of cases within subregion i .

For rare diseases such as leukemia, the null hypothesis is equivalent to: $H_0 : C_i (i = 1, \dots, I)$ are independent Poisson random variables with $E[C_i] = \lambda n_i$. λ is the per person rate. The test types differ in their alternative hypotheses. Tests with the alternative $H_a : \text{not } H_0$ are called *general tests*. Tests

Table 1. TCE waste sites and the counties in which they are located

Site number	Name	County
1	Monarch Chemicals	Broome
2	IBM Endicott	Broome
3	Singer	Broome
4	Nesco	Broome
5	GE Auburn	Cayuga
6	Solvent Savers	Chenango
7	Smith Corona	Cortland
8	Victory Plaza	Tioga
9	IBM Owego	Tioga
10	Hadco	Tioga
11	Morse Chain	Tompkins

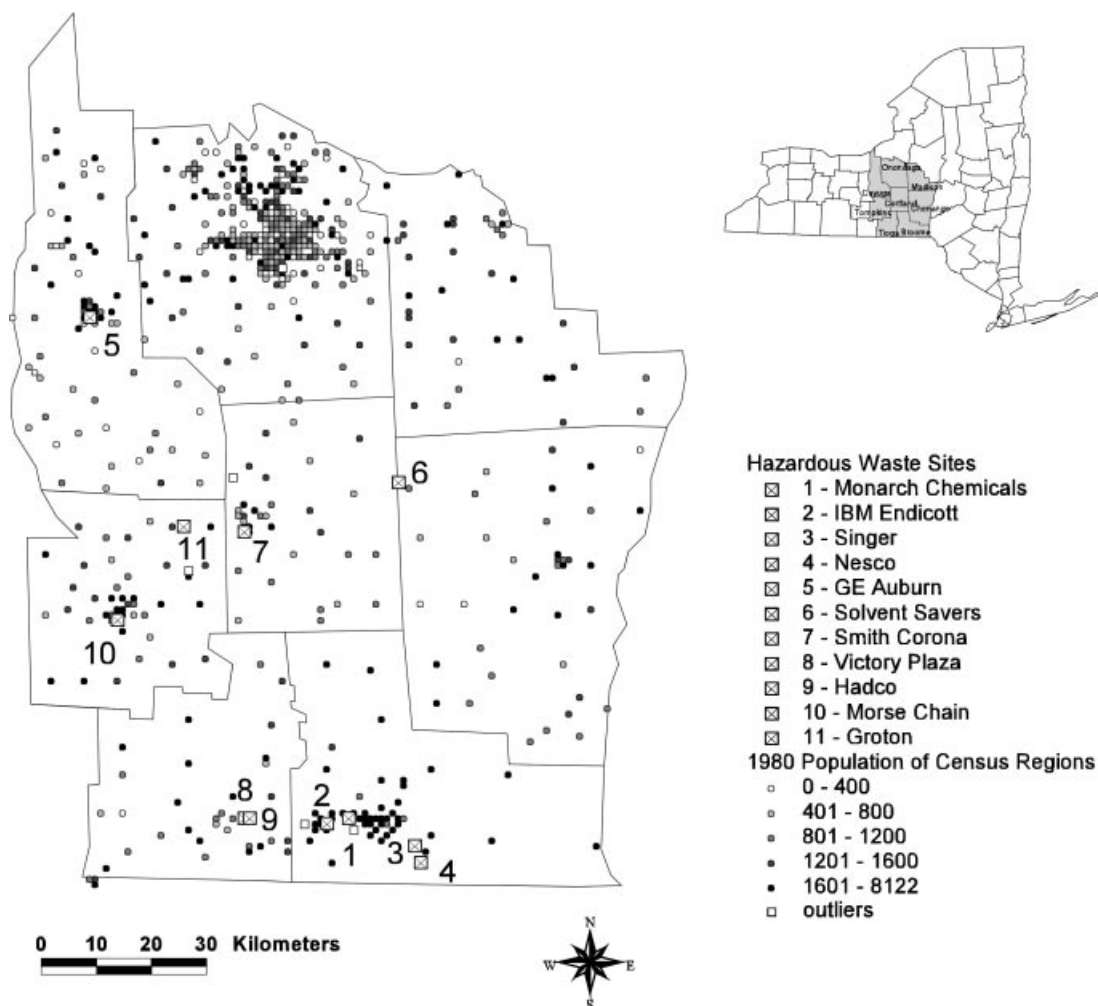


Figure 1. Population centroids and TCE waste sites. The population centroids are shaded to denote the population of the corresponding census region

with a more specific alternative such as $H_a : E[C_i] = \lambda n_i(1 + d_i)$ (d_i is the inverse of the distance from location i to a suspected source) are called *focused tests*.

Waller *et al.* (1992) built on Turnbull *et al.* (1990), which applied the following general tests (test without the foci) to the New York data: the GAM (geographical analysis machine) method by Openshaw *et al.* (1988), U -statistic of Whittemore *et al.* (1987), and the cluster evaluation permutation procedure (CEPP) developed by Turnbull *et al.* (1990). Using these tests, the evidence of clustering is weak, although there is some suggestive clustering in Cortland, Broome and Cayuga Counties.

Waller *et al.* (1992) then used the 11 TCE-contaminated waste sites as the putative sources of hazard (the foci), and applied several focused tests to the data: a focused version of the method of Besag and Newell (1991), a test by Waller (1992) and a focused test by Stone (1988).

While Besag and Newell's test does not indicate clustering around any of the 11 waste sites, Stone's procedure yields significant values at the Monarch Chemical and IBM Endicott sites in Broome County. However, neither the Stone test nor the test by Besag and Newell find statistically significant clusters when the multiplicity of tests is taken into account. In contrast, Waller's focused test (1992) gives several significant results, again showing Monarch Chemical as the focus of the most likely cluster.

1.3. Literature review

Since 1992, these data have been reanalyzed in several published papers. Waller and Turnbull (1993) used the data while discussing the effects of scale on testing for disease clustering. Kulldorff and Nagarwalla (1995) applied their new general cluster detection methodology to these data and found probable clustering in Broome county. While there may have been clustering in other counties, their methodology was not designed to detect more than one cluster. Waller (1996) defined the power functions for common focused tests and illustrated his results with these data. More recently, Gangnon and Clayton (1998) and Ghosh *et al.* (1999) applied Bayesian cluster analysis methods to the data. Gangnon and Clayton's method found probable clustering in Broome, Cortland and Onondaga Counties. Ghosh *et al.*'s focused method was applied to all the sites simultaneously and found suggestive but inconclusive evidence of clustering. Rogerson (1999) developed and applied chi-squared focused and general tests to the data. He also found evidence of clustering around Monarch Chemical.

Note that none of the above studies used covariates, although most of the methods can accommodate covariate information, and all the authors noted that their analyses were potentially confounded. Of course, many of those papers were using the data to exemplify their new methodologies and epidemiological inference was not their primary concern. Recently, Waller and McMaster (1997) analyzed the Broome County subset of the data using counts which were externally age standardized using the National Cancer Institute's Cancer Surveillance, Epidemiology and End Results (SEER) data (Horn *et al.*, 1984). Compared to an unstandardized analysis, they found that standardization increased the statistical significance of the association between the outcome and proximity to Monarch Chemical. We are aware of no other published analysis that includes any covariate information.

In this article, we analyze these data taking covariates into consideration. Our findings, reported in Section 2.5, suggest that the effect of being near Monarch Chemical and G.E. Auburn are at least partially confounded by age and occupation. We find a significantly increased leukemia rate associated with living in an area with a high percentage of manufacturing employment. We also identify proximity to the Smith Corona site to be significantly associated with increased leukemia counts after controlling for other covariates. The Smith Corona site was not previously identified as a possibly dangerous site. Note that our results are still possibly confounded by unmeasured covariates.

2. COVARIATE ADJUSTED TCE SITE/LEUKEMIA RELATIONSHIP

This section describes our analysis of the NYS TCE data, incorporating covariates derived from 1980 US census data. It consists of three parts. First, the additional covariate data are described. We then discuss our model-building strategy. Finally, we discuss the fitted models and conclusions based on them.

2.1. Data

We augmented Waller's 1992 data with demographic data from the 1980 census. Data at the block group level were available from two sources: Summary Tape Files 1A and 3A (STF1A and STF3A). Data from STF1A included information on the percentage of respondents in each block group of a given race, age and house value, and whether the block group was urban or rural. STF1A had data for each of Waller *et al.*'s 790 census regions. STF3A recorded block group level information about education level, employment (both industry and job type), income and source of drinking water. Note that the source of drinking water variable is a binary variable listing public sources versus private sources and not a list of particular wells.

STF3A was missing data for 182 block groups which included all of Tompkins and Tioga counties and 21 percent of Onondaga County, a spatially contiguous portion of upstate New York. This is a systematic missingness. It is due to the sampling strategy of the Census Bureau, which is independent of health outcomes. Because of the spatial aspect of the missing demographic data, use of data imputation seemed invalid. For instance, Tompkins County contains Cornell University and is possibly quite different from the other counties in the area. As a result, we did not pursue missing data strategies. Instead, we conducted two analyses: one on the complete data using only STF1A information and one on the available data for both STF1A and STF3A. The block group populations in both files matched Waller *et al.*'s Waller *et al.* (1992) data exactly.

2.2. Model building strategy

With leukemia rates per block group as the response, our model building strategy had three steps. First we used a Box-Cox transformation to normalize the data and chose variables for the model using all subsets linear regression. As expected, the Box-Cox procedure suggested a log transform for the leukemia rates. The variables selected (using this procedure and others described below) are listed in Table 2. Since we searched through the data to find significant variables, we need to adjust our significance levels appropriately.

In the second step we selected a distance function to relate the TCE sites to the cases. Partial leverage plots (Figure 2) suggested using the inverse of the distance to a site in kilometers if the census group's centroid is within 20 km of the site and zero otherwise. While such a distance function could result in infinite or poorly scaled variables, for this dataset it does not. All values are less than 3.1 km^{-1} and most are less than 1 km^{-1} . The partial leverage plots for the STF3A are in Figure 2. STF1A's partial leverage plots are similar.

Once we had the set of candidate variables and the distances, our third step was to pare the list down by removing variables which were highly intercorrelated. As the map (Figure 1) shows, several sites in the southern part of our region of interest are close together due to an 'industrial corridor' along the Susquehanna River. To prevent collinearity of distance measures, we removed IBM-Endicott, Singer, Victory Plaza, IBM-Owego, Nesco and Hadco. Since it is near the center of that region and was referred to in Waller (1992), Monarch Chemical was retained as the site from that area. The variable 'Near Monarch' is a proxy for proximity to all the sites listed above.

Based on correlations between and among the sites and the other covariates, we retained the variables listed in Table 2.

2.3. The model

We fitted generalized linear models to these data. We used a log link for the mean and a variance function which is proportional to the mean. The model corresponds to Poisson regression with a

Table 2. Information included in 1980 census summary tape files 1A and 3A (STF1A and STF3A) and which variables we include in our analyses

STF1A All subsets selected	In analysis	STF3A All subsets selected	In analysis
% age over 60	Yes		Yes
Urban indicator	No		No
% white	No		No
% black	No		No
% house value ≤ \$10K	No		No
% house value \$15-20K	No		No
% house value \$20-25K	No		No
% house value \$30-35K	No		No
		% on public water	Yes
		% protection/service jobs	Yes
		% other service jobs	No
		% technician jobs	Yes
		% farming/fishing jobs	No
		% precision repair jobs	Yes
		% total manufacturing jobs	Yes
Near Monarch	Yes		Yes
Near IBM-End	No		No
Near Singer	No		No
Near Nesco	No		No
Near GE	Yes		Yes
Near Solvent	Yes		Yes
Near Smith	Yes		Yes
Near Victory	No		No
Near IBM-Owego	No		No
Near Hadco	No		No
Near Morse	Yes		No

relaxation of the requirement that the variance equals the mean. (See Chapter 10 of McCullagh and Nelder, 1989.) Census blocks with populations less than four hundred were not used in the analysis, and some blocks with outlying observations were also removed. The random variables representing the incident leukemia rates are assumed to be independent across block groups.

One advantage of these models over the cluster analysis procedures discussed in Section 1.2 is that these models yield quantitative risk and rate estimates in addition to finding clusters. As pointed out by Gangnon and Clayton (1998), modeling instead of looking for clusters also has the benefit of forcing the analyst to go through the meaningful exercise of explicitly formulating, debugging and testing the model.

The fitted models are below:

STF1A model:

$$\begin{aligned} \log E[\text{Cases}_i / \text{Pop}_i] = & \beta_0 + \beta_1 \text{Age over } 60_i + \beta_2 \text{Near Monarch}_i \\ & + \beta_3 \text{Near GE}_i + \beta_4 \text{Near Solvent}_i + \beta_5 \text{Near Smith Corona}_i \\ & + \beta_6 \text{Near Morse}_i \end{aligned}$$

$$\text{Var}[\text{Cases}_i] = \phi E[\text{Cases}_i]$$

$$i = 1, \dots, 790 \text{ census blocks}$$

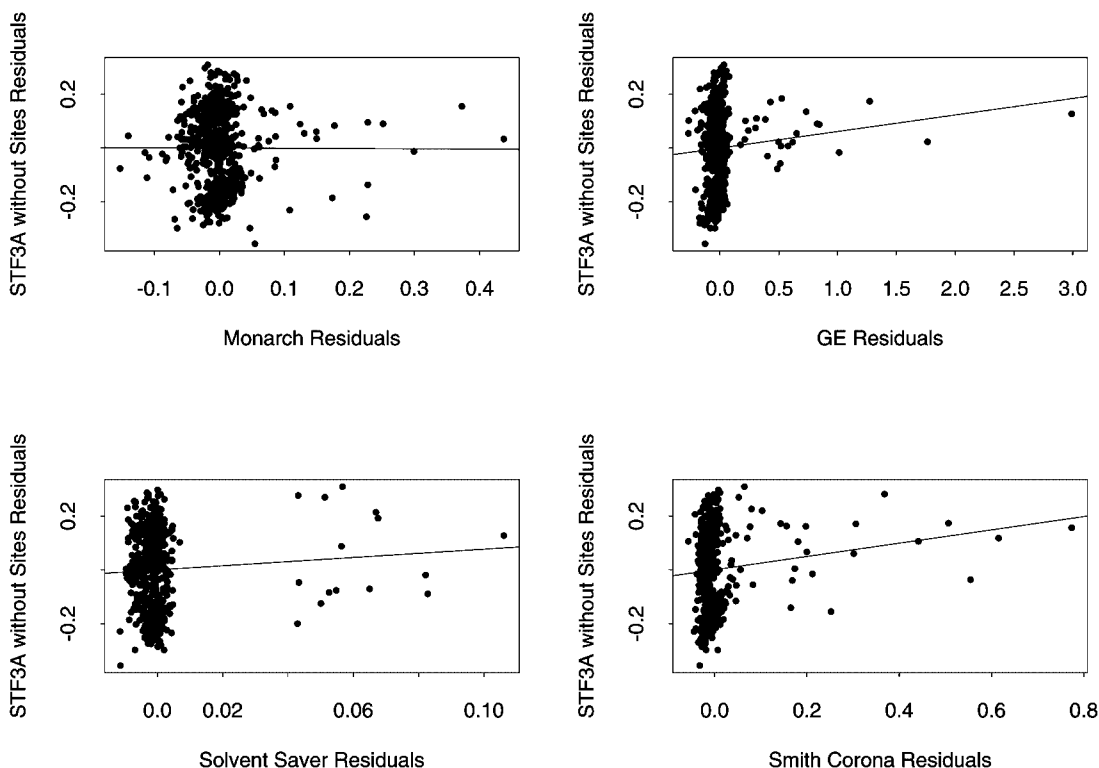


Figure 2. Partial leverage plots for $1/(km\ to\ the\ TCE\ waste\ sites)$. These partial leverage plots are from a linear version of the STF3A model with the square root of the case rate as the response. Partial leverage plots for STF1A's model look similar

STF3A model:

$$\begin{aligned} \log E[Cases_i/Pop_i] = & \beta_0 + \beta_1 Age\ Over\ 60_i + \beta_2 Public\ Water_i + \beta_3 Protect\ Service_i \\ & + \beta_4 Precision\ Repair_i + \beta_5 Technician_i + \beta_6 Total\ Manufacturing_i \\ & + \beta_7 Near\ Monarch_i + \beta_8 Near\ GE_i \\ & + \beta_9 Near\ Solvent_i + \beta_{10} Near\ Smith\ Corona_i \end{aligned}$$

$$\begin{aligned} Var[Cases_i] = & \phi E[Cases_i] \\ & i = 1, \dots, 608\ census\ blocks \end{aligned}$$

2.4. Analysis

The models presented above yield two pieces of information: coefficient estimates and analysis of deviance tables. The coefficient estimates are in Table 3 and the analyses of deviance are in Table 4.

In generalized linear models, a variable's contribution to deviance divided by total deviance with no variables in the model (null deviance) can be considered to be the percentage of fit attributable to that variable. When the design is not orthogonal or the data are not normally distributed, the deviance depends on the order in which the variables were entered into the model. To give a sense of the

Table 3. Summary of the estimated coefficients and *p*-values based on *t*-statistics. Note that the *p*-values are not adjusted for the variable selection procedure

Variable	STF1A coeff.	<i>t</i> -stat.	<i>p</i> -value	STF3A coeff.	<i>t</i> -stat.	<i>p</i> -value
Age Over 60	4.1	7.6	0.00	2.6	3.71	0.00
Public Water				0.5	1.51	0.07
Protect Service				8.0	1.21	0.11
Precision Repair				-10.6	-3.91	0.00
Technicians				-8.6	-1.68	0.05
Total Manufacturing				3.1	2.18	0.01
Near Monarch	1.0	1.39	0.08	0.7	1.00	0.16
Near GE	0.2	1.32	0.09	0.2	0.89	0.19
Near Solvent	0.8	0.36	0.36	1.0	0.50	0.31
Near Smith Corona	1.9	4.29	0.00	1.6	3.44	0.00
Near Morse	-0.1	-0.38	0.35			

Table 4. The change in deviance associated with each covariate. We list the change in deviance when the covariate is entered into the model first and last. Note that these values do not necessarily bound all the possible values of deviance for all possible orders of variable addition

Variable	STF1A first	last	STF3A first	last
Age Over 60	0.0474	0.0429	0.0279	0.0107
Public Water			0.0167	0.0018
Protect Service			0.0034	0.0011
Precision Repair			0.0203	0.0126
Technician			0.0008	0.0023
Total Manufacturing			0.0006	0.0037
Near. Monarch	0.0039	0.0015	0.0040	0.0007
Near GE	0.0026	0.0012	0.0012	0.0006
Near Solvent	0.0000	0.0001	0.0000	0.0002
Near Smith Corona	0.0100	0.0111	0.0100	0.0074
Near Morse	0.0005	0.0001		
Null Deviance	0.5781		0.4274	

contribution of fit for each variable in our model, we list two deviances for each, one when it is included in the model first and one when it is put in last. Note that these values *do not* necessarily bound all the possible values of deviance for all possible orders of variable addition.

Note that, for both models, the deviance attributable to distance to the TCE sites is a very small percentage of the total deviance. While there are tests of significance based on deviance, we base our tests on the *t*-statistics since we have such a large sample size.

2.5. Interpretation

Due to missing data, non-orthogonal design, and use of all subsets regression, interpreting the results requires extra care. Note that significant associations between sites and disease do not establish causality. See Section 2.8 for more discussion of this issue.

Table 5. Summary of how the parameters associated with the TCE waste sites change as demographic covariates are added to the model. Note that Near Smith Corona is the only covariate that remains nominally statistically significant after age and occupation are added from STF1A and STF3A. Near Morse is not included in the STF3A analysis since it is in Tompkins county and that county's data is not included in STF3A

Covariate	Base case est./t-value	STF1A est./t-value	STF3A est./t-value
Near Monarch	1.70/2.42	0.98/1.39	0.72/1.00
Near GE	0.38/2.07	0.24/1.32	0.18/0.89
Near Solvent	0.08/0.03	0.78/0.36	1.03/0.50
Near Smith Corona	1.89/3.96	1.94/4.29	1.60/3.44
Near Morse	-0.16/-0.50	-0.11/-0.38	

Consider the results for STF1A first. At a nominal 10 per cent significance level, the variables age, near Monarch, near GE and near Smith Corona are all associated with increased incident leukemia rates. Further, age is identified as a risk factor for leukemia, as expected (Linnet, 1985). The sites identified as statistically significant are consistent with the sites identified by age unadjusted studies cited in Sections 1.2 and 1.3. Our age adjustment decreased the estimated relationship (see Table 5). In Table 4, note that the deviance associated with the variable Near Monarch varies greatly depending on whether or not other covariates are used. This suggests that Near Monarch's relation to the response is strongly influenced by the other variables in this analysis.

Consider the STF3A analysis next. Age remains significant. There is a suggestion of increased risks associated with drinking public water and working in 'manufacturing'. A high percentage of the population working as technicians or in a jobs classified as precision repair are associated with lower rates. We do not have an explanation for this. In a departure from previous analyses, Table 3 shows that the only site variable associated with increased leukemia rates after controlling for other covariates is Near Smith Corona. We investigated and found no evidence of significant interactions between age and any of the employment variables.

Tables 3–5, and correlation calculations, suggest that the previous results about the G.E. and Monarch sites were confounded by age and occupation. More specifically, the percentage of residents over 60 years old, the percentage working in jobs classified as precision repair and manufacturing, and drinking public water are related to both the site proximity and leukemia counts. Of those variables, only removing the predictor PublicWater from the analysis had little effect on the estimates in Table 5. Hence we conclude that the clustering that was attributed to site proximity might actually be due to clustering of residents in certain occupations and population age.

A natural question is whether occupation is a proxy for proximity to Monarch and G.E. This is unlikely because both the areas near Monarch and G.E. are relatively industrialized, and the employees of those two companies probably made up a small part of the total manufacturing employment in their respective areas. Specific information about the number of people employed in manufacturing at those two company sites during the time of interest is not available.

Results differ between the STF1A and STF3A models. It is probable that some of the change in the results from STF1A to STF3A is due to a lack of power. More specifically, since STF3A has both more covariates and fewer observations, we would expect *p*-values to increase. The difference in results for variable near GE may fall into this category. On the other hand, while near Monarch's *p*-value also increases, that change is accompanied by a large change in its coefficient. That suggests that the effects of being near Monarch is confounded with other variables in STF3A. It is worth pointing out that only the Morse site had more than 5 per cent of the observations within 20 km missing.

Table 6. Summary of the estimated rate ratios from each model. Note that although there is a large rate associated with Protect Service, that rate is not statistically significantly different from 1

Variable	STF1A rate ratio	STF3A rate ratio
Age Over 60 [¶]	60.3	13.5
Public Water		1.6
Protect Service		2981.0
Precision Repair [§]		2.5×10^5
Technicians [§]		1.8×10^3
Total Manufacturing [§]		22.2
Near Monarch [‡]	2.7	2.0
Near GE [‡]	1.2	1.2
Near Solvent	2.2	2.7
Near Smith Corona [¶]	6.7	5.0
Near Morse	0.9	

[‡]Nominally significant at 10% level in STF1A analysis only.

[§]Nominally significant at 10% level in STF3A analysis only.

[¶]Nominally significant at 10% level in both STF1A and STF3A.

2.6. Statistical significance and public health significance

One measure of public health significance is the rate ratio:

$$\frac{\text{incidence rate for exposed}}{\text{incidence rate for unexposed}}$$

(Kelsey *et al.*, 1996). While this measure is for a dichotomous exposure (exposed versus non-exposed), under the Poisson model, the parameter estimates refer to the effect of a one unit change in the exposure on the log of the rate of the outcome. As a result, the anti-log of each parameter estimate is a 'rate ratio'. Table 6 summarizes the rate ratios associated with each variable for the STF1A and STF3A models.

From Table 6 it can be seen that, of the nominally significant variables, 'Age Over 60' and 'Total Manufacturing' seem to have the most effect on the rate. The point estimates of the rate ratios associated with the TCE sites are much smaller. Confidence intervals around these point estimates are consistent with this interpretation.

2.7. Other approaches

Other models besides generalized linear models are feasible for these data. We outline some of these below.

Cases whose addresses were unknown were allocated proportionally across the block groups in areas of likely residence. This suggests that an appropriate model should take into account the multiple resolutions at which the data were collected. Future theoretical work will make this idea more specific.

An aspect of these data which we did not consider is that, even after adjusting for the covariates, the number of cases per census block group may be spatially correlated. A paper by Ghosh, *et al.* (1999) suggests a way to address that aspect of the data using hierarchical Bayesian models (Besag *et al.*, 1991, 1995).

Another way to relax the model assumptions would be to consider Generalized Additive Models (Hastie and Tibshirani, 1990). An interesting future project would be to combine this approach with the hierarchical Bayesian approach cited above.

Finally, another fairly simple modeling approach would have been to conduct a cluster analysis on the residuals from the two regression models without the inverse distances to the TCE sites. We tried this by applying a version of the score test designed by Waller (1992), modified to account for the possible negative values of the residuals. No evidence of significant clustering around any of the 11 foci was discovered.

In addition to different modeling approaches, it would be useful to investigate additional data sources as well. Newer case and covariate data would be more relevant. Also, the exposure variables could be improved. While TCE storage sites are one source of industrial exposure into the environment, we suspect that there are many others also. The Cornell University Geospatial Information Repository contains Graphical Information System (GIS) overlays containing spatial information about annually reported industrial chemical releases in New York State. Further exploration of how GIS capabilities could be used in a study like this also would be interesting.

2.8. *Causation versus association*

In addition to the statistical concerns discussed in the previous section, the primary weakness with this analysis lies in the fact that it is an ecological study. See Chapter 23 of Rothman and Greenland (1998), for instance. Since the cases and covariates are aggregated to the block group (or tract) level, we do not know if the covariates actually apply to the cases. Studies such as these certainly can suggest a causal relationship, but they provide much weaker evidence than a cohort or case control study, for instance.

A second factor which makes it difficult to establish causality for this study is that case incidence and site exposures are measured concurrently. Better data would consider the induction period of leukemia and measure exposure before incidence and the potential problem of population mobility into and out of the area of the study during that period.

Finally, it is important to emphasize that this study, like other 'cluster analysis studies', is based on non-experimental data. As a result, it is always going to be vulnerable to unknown and observed confounding covariates which potentially invalidate causal inferences. As an anonymous referee pointed out, these studies 'inherently lack the randomization mechanisms that would help to average out the extraneous confounding factors'. A statistical review paper discussing confounding and non-experimental data can be found in Greenland *et al.* (1999).

3. CONCLUSION

3.1. *Summary of results*

Our analysis improves on previous analyses of these data by taking additional relevant covariates (age, occupation and water source) into account. While previous authors stated the desirability of adjusting for covariates, only Waller and McMaster (1997) added the available census data.

After controlling for those covariates, we found a relationship between being near the Smith Corona site and an elevated leukemia rate in 1978–1982. While previous analyses had found relationships between elevated rates and other sites, our analysis suggests that those results were confounded by occupation, industry, and age.

3.2. Current status of the sites

Monarch Chemical, GE Auburn and Smith Corona were the three sites that proved most interesting in our analyses. Monarch Chemical, which is currently on the national Comprehensive Environmental, Response, Compensation and Liability (CERLIS) Hazardous Waste Site list, has participated in clean-up efforts. In 1982 two wells in the Vestal Water Supply system near Monarch were found to be contaminated with high levels of TCE; the wells were closed until a treatment system was constructed, and were reopened in 1988. New York State took legal actions against Monarch Chemical and other potentially responsible parties, and an agreement was signed in 1985. As part of the signed agreement, the potentially responsible parties paid to have 42 tons of contaminated soil removed. Levels of contaminants in untreated ground-water have since declined to levels approaching drinking water standards, and the site remediation was considered complete in September 1998 (United States Environmental Protection Agency).

The site at GE Auburn is listed in the national No Further Remedial Action Planned (NFRAP) site list. The only actions listed for this site were an Initial Discovery in June 1981 and a Preliminary Assessment in June 1987. The Smith Corona site in Cortland is not listed on any national waste site lists; the authors are not aware of what cleanup actions, if any, have been taken.

3.3. Environmental epidemiological studies and public policy

As concern over cancer rises in our society, it becomes ever more important to address the societal impact of studies of disease clusters. Both the policy and legal impacts can directly affect individuals.

Cancer is likely to be associated with increased perceptions of risk for several reasons:

- children may be directly affected;
- individuals perceive that the disease cannot be avoided or controlled;
- it is associated with feelings of dread;
- the prevalent perception is that individuals are exposed to carcinogenic compounds without their knowledge or consent.

For example, the town of Woburn, Massachusetts has become the focus of much publicity beginning in 1982, when families of children with leukemia filed suit against two local companies for medical damages that they alleged resulted from the contamination of the local water wells by the companies (Lagakos *et al.*, 1986; Fienberg and Kaye, 1991). The case is now the subject of a book, *A Civil Action* (Harr, 1995), and a major motion picture of the same name. The Woburn case, and others like it (civil mass tort cases for environmental contamination and the damage allegedly caused to individuals as a result) rely heavily upon statistical evidence. The courts will be increasingly expected to instruct juries regarding the use of statistical evidence.

As cases such as Woburn become more widely known, these concerns will continue to increase, as will reports of potential clusters (Gawande, 1999). Cancer cluster studies can place a heavy financial and time burden on agencies, especially when searching for site specific incidence. The combination of the rare disease rate and the small number of people living close to any one facility can create difficulties in even the best-designed studies. Non-significant findings may not lay concerns to rest in the eyes of the affected communities.

There are also political and policy ramifications. Agencies, such as environmental protection agencies, have individually tailored responsibilities and overriding management goals. However, they must also be responsive to the concerns of the citizenry and of local, state and national political decision-makers in order to survive politically and establish a support base within the community

(Jones, 1984). These requirements can cause complications when dealing with cancer cluster studies for reasons that will be outlined below.

There are three reasons that 'non-significant' findings may still have policy implications:

- poor public understanding of the statistical and scientific issues;
- differing views of 'acceptable' risk among the affected communities and the policy-making agencies;
- disproportionate distribution of potential hazards in low-income or otherwise disadvantaged communities.

Agencies and individuals involved in communicating study results to the public must be aware of these problems and the potential for the findings to become part of our legal, political and policy environments.

Whenever possible, a brief synopsis of results should be available to the public in easy-to-understand language. Increased focus on methods of risk communication specifically tailored to cancer cluster incidence may be beneficial for improved agency communication with the public before, during and after cancer cluster studies. Agencies may be able to use these methods to effectively communicate with the public about the *need* for a study in the first place, which may mitigate some of the strain placed on agency budgets by the increased demand for cluster studies. A well thought out procedure to address cancer cluster reports that was undertaken by the Minnesota Department of Public Health is described in Bender *et al.* (1990).

ACKNOWLEDGEMENT

This work was supported by NIEHS training grant number ES07261 at Cornell University. We thank the editor and anonymous referees for helpful comments.

REFERENCES

- Bender A, Williams A, Johnson R, Jagger, H. 1990. Appropriate public health responses to clusters: the art of being responsibly responsive. *American Journal of Epidemiology* **132**(4): S48–S52.
- Besag J, Green P, Higdon D, Mengersen K. 1995. Bayesian computation and stochastic systems (with discussion). *Statistical Science* **10**: 3–66.
- Besag J, Newell J. 1991. The detection of clusters in rare diseases. *Journal of the Royal Statistical Society A* **154**(1): 143–155.
- Besag J, York J, Mollie A. 1991. Bayesian image restoration with two applications to spatial statistics. *Annals of the Institute of Statistical Mathematics* **43**: 1–59.
- Fienberg S, Kaye DH. 1991. Legal and statistical aspects of some mysterious clusters. *Journal of the Royal Statistical Society A* **154**(1): 61–74.
- Gangnon RE, Clayton MK. 1998. Bayesian spatial disease clustering: an application. *Technical Report #132*. Department of Biostatistics, University of Wisconsin–Madison.
- Gawande A. 1999. The cancer-cluster myth. *The New Yorker* 9 February: 34–37.
- Ghosh M, Natarajan K, Waller L, Kim D. 1999. Hierarchical Bayes GLMs for the analysis of spatial data: an application to disease mapping. *Journal of Statistical Planning and Inference* **75**: 305–318.
- Greenland S, Robins J, Pearl J. Confounding and collapsibility in causal inference. *Statistical Science* 1999; **14**: 29–46.
- Harr J. 1995. *A Civil Action*. Random House: NY.
- Hastie T, Tibshirani R. *Generalized Additive Models*. Chapman & Hall: London, 1990.
- Horn JW, Asire AJ, Young JL Jr, Pollack ES (eds). 1984. *SEER Program: Cancer Incidence and Mortality in the United States 1973–1981*. Bethesda, MD, National Institute of Health Publication No. 85-1837.
- Iwano E. 1989. A comparison of cluster detection procedures. *MS thesis*. Department of Operations Research and Industrial Engineering, Cornell University.
- Jones CO. 1984. *it/ An introduction to the study of public policy* (3rd edn). Harcourt Brace: Fort Worth.

- Kelsey J, Whittemore A, Evans A, Thompson D. 1996. *Methods in Observational Epidemiology*. Oxford University Press: New York.
- Kimbrough R, Mitchell F, Houk V. 1985. Trichloroethylene: an update. *Journal of Toxicology and Environmental Health* **15**: 369–383.
- Kulldorff M, Nagarwalla N. 1995. Spatial disease clusters: detection and inference. *Statistics in Medicine* **14**: 799–810.
- Lagakos SW, Wessen BJ, Zelen M. 1986. An analysis of contaminated well water and health effects in Woburn, Massachusetts (with discussion). *Journal of the American Statistical Association* **82**: 583–596.
- Lange N, Ryan L, Billard L, Brillinger D, Conquest L, Greenhouse J (eds). 1994. *Case Studies in Biometry*. Wiley: New York.
- Linet MS. *The Leukemias: Epidemiologic Aspects*. Oxford University Press: New York, 1985.
- McCullagh P, Nelder JA. 1989. *Generalized Linear Models (2nd edn.)*. Chapman & Hall: London.
- Openshaw S, Craft AW, Charlton M, Birch JM. 1988. Investigation of leukemia clusters by use of a geographical analysis machine. *Lancet* 272–273.
- Rogerson P. 1999. The detection of clusters using a spatial version of the chi-square goodness-of-fit statistic. *Geographical Analysis* **31**(1): 130–147.
- Rothman K, Greenland S. 1998. *Modern Epidemiology*. Lippincott–Raven: Philadelphia, PA.
- Stone R. 1988. Investigations of excess environmental risks around putative sources: statistical problems and proposed tests. *Statistics in Medicine* **7**: 649–660.
- Turnbull B, Iwano E, Burnett W, Howe H, Clark L. 1990. Monitoring for clusters of disease: application to leukemia incidence in upstate New York. *American Journal of Epidemiology* **132**(4): S136–S143.
- United States Environmental Protection Agency. 1999. Superfund National Priority Site Fact Sheet, Vestal Water Supply Well 42. www.epa.gov/r02earth/superfund/sitesum/0202152c.htm
- United States Environmental Protection Agency. 1999. Superfund Archive Sites: No Further Remedial Action Planned (NFRAP), General Electric/Auburn Plant. www.epa.gov/oerrpage/superfund/sites/arcsites/reg02/a0201449.htm
- Waller L. *PhD Thesis*. 1992. Department of Operations Research and Industrial Engineering, Cornell University.
- Waller L. 1996. Statistical power and design of focused clustering studies. *Statistics in Medicine* **15**: 765–782.
- Waller L, McMaster R. 1997. Incorporating indirect standardization in tests for disease clustering in a GIS environment. *Geographical Systems* **44**(4): 327–342.
- Waller L, Turnbull B. 1993. The effects of scale on tests for disease clustering. *Statistics in Medicine* **12**: 1869–1884.
- Waller L, Turnbull B, Clark L, Nasca P. 1992. Chronic disease surveillance and testing of clustering of disease and exposure: application to leukemia incidence in TCE-contaminated dumpsites in upstate New York. *Environmetrics* **3**: 281–300.
- Whittemore A, Friend N, Brown B Jr, Holly E. 1987. A test to detect clusters of disease. *Biometrika* **74**(3): 631–635.