APPLICATION OF A TIME-SPACE CLUSTERING METHODOLOGY
TO THE ASSESSMENT OF ACUTE ENVIRONMENTAL EFFECTS
ON RESPIRATORY ILLNESSES

by

Inge F. Goldstein
Columbia University School of Public Health

and

Jack Cuzick
Columbia University, Department of Mathematical Statistics

June, 1978

STUDY ON STATISTICS AND ENVIRONMENTAL
FACTORS IN HEALTH

PREPARED UNDER SUPPORT TO SIMS FROM

DEPARTMENT OF ENERGY (DOE)
ENVIRONMENTAL PROTECTION AGENCY (EPA)
NATIONAL SCIENCE FOUNDATION (NSF)
ROCKEFELLER FOUNDATION
SLOAN FOUNDATION

COLUMBIA UNIVERSITY
NEW YORK, NEW YORK

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

APPLICATION OF A TIME-SPACE CLUSTERING METHODOLOGY
TO THE ASSESSMENT OF ACUTE ENVIRONMENTAL EFFECTS ON RESPIRATORY ILLNESSES*

Inge F. Goldstein
Columbia University School of Public Health
600 West 168th Street, New York, N.Y. 10032

Jack Cuzick
Columbia University
Department of Mathematical Statistics

Abstract

A new methodology is proposed for the identification of environmental events of health significance. Health indices measured on a daily basis at various locations in a single geographical area are collected over time. First, the daily variations are examined to determine whether they reflect purely random variations or whether there are days on which there are extreme variations not plausibly explicable as random events. After such days are identified, the question of whether they occur only at a single location within the larger geographical area at one time, or whether they occur simultaneously at more than one location is investigated. Tests of statistical significance for both temporal and spatial clustering are proposed.

The methodology is applied to daily hospital emergency room visits for various respiratory complaints to several New York City hospitals situated in two geographically separated districts which however have populations of similar socio-economic and ethnic composition.

* This work was supported by research grants from the National Institute of Environmental Health Sciences, Grant #00899, and by SIMS Study on Statistics and Environmental Factors in Health under the auspices of the SIAM Institute for Mathematics and Society.

APPLICATION OF A TIME-SPACE CLUSTERING METHODOLOGY
TO THE ASSESSMENT OF ACUTE ENVIRONMENTAL EFFECTS ON RESPIRATORY ILLNESSES

Inge F. Goldstein and Jack Cuzick

Introduction

The relationship between environmental factors and their possible acute effects on health is extremely complex. We have discussed in earlier papers, [1-6] some of the problems that arise when these relationships are investigated in epidemiologic studies in particular, the limitations and drawback of many studies that make use of linear regression analyses in which measured pollutants or other environmental factors are regressed on mortality or morbidity occurring in specified short periods of time (such as hours or days). At best, only a few of the many possible environmental agents that people are exposed to are presently measured. In addition, many of these environmental agents interact with each other or may have synergistic effects on health. The accuracy with which some of the most commonly monitored pollutants is measured is extremely low (less than 100% at the present ambient levels of sulfur dioxide, for example). Further, the extent to which these measurements represent pollution in the geographic areas surrounding the site of measurement is usually unknown, thus adding to the problems of specification of the independent or exposure variable.

We have been engaged for some time in a search for exploratory methods of identifying possible environmental "events" having harmful effects that do not at the start require that specific pollutants be

identified as possible causal agents. Instead, we search first for temporo-spatial clustering of morbidity or mortality, and only when such clusters are identified do we search for causal environmental factors. We have published a study [7] in which we identified a small number of days on which visits to hospital emergency rooms for asthma were unusually high and we then showed that such days tended to occur simultaneously at emergency rooms all over the New York metropolitan area. This in turn implies that whatever environmental factor was responsible appeared over the entire metropolitan area at the same time.

In this paper we describe a somewhat different methodology which can also be used for the purpose of identifying environmental events of health significance. Again, we focus on health indices measured on a daily basis, and ask, first: how they vary over time, and second, how the daily variations compare at different locations in the city. We can thus identify time clustering in the occurrence of specific illness, and also distinguish temporal variations that are local to specific geographic areas of the city from those that occur globally over the whole city. Once the temporo-spatial patterns of the health indices are discovered, a search for environmental etiologic agents can be limited to those whose temporo-spatial pattern is concordant with the observed health outcome.

While the methods of analysis we describe here are applied for illustrative purposes to respiratory complaints, they can obviously be applied to many other conditions as well. Information gained about

disease patterns could also be of great benefit in the design of aerometric and other networks whose function is to monitor possible environmental factors associated with disease, and help answer such questions as: How many monitoring stations should a certain city have? What pollutants should be measured? How often? etc.

Method and Procedure

Daily visits to emergency rooms at three major metropolitan hospitals in two inner-city areas of New York City were collected for a period of 3 years (1969-1971). The map (Figure 1) shows the location of the hospitals; two of them, Kings County and Cumberland, are situated in one area of Brooklyn, whereas the third hospital, Harlem, is located in upper Manhattan.

The districts served by the three hospitals are low-income areas and the population rarely [8] uses private physicians for its medical care; it is mostly non-white, and most of those using the hospitals live in the health areas immediately surrounding them. All respiratory complaints were coded.* After coding, all respiratory visits were allocated to three major diagnostic subdivisions, asthma (which consisted of all asthmatic diagnoses), upper respiratory visits, and lower respiratory visits (which consisted mainly of bronchitis, pneumonia and influenza). The visits to each emergency room were examined separately first in order to get an idea of the overall temporal variability; then they were studied in pairs to investigate how much of the temporal variation is common to the pairs.

* Detailed directions for coding are available from

One Sample Case

Each individual hospital was first examined to determine if the variations in the daily numbers of visits arise solely from random fluctuations, or if they are associated with environmental conditions which influence the state of health differently on certain days. If the population at risk is large and stable, and the probability that any one person will show up at the emergency room on any given day is small and dependent only on his susceptibility to the disease and independent of environmental factors that change over time-scales of the order of a day or longer, then the number of events on day i , written N_i , is approximately a Poisson variable. These assumptions will be invalid of course when environmental agents to which many individuals are susceptible vary with time, and also if the disease which brings them to the emergency room is highly contagious. Our method will detect temporo-spatial clustering, but a separate determination must be made as to whether these clusters are due to a localized environmental agent or to contagion.

Let us denote the expected value of N_i by λ_i . The hypothesis that there is no temporal variation implies that $\lambda_i \equiv \lambda$ is the same for all days (although unknown). The d statistic [9]

$$(1) \quad d = \frac{\sum_{i=1}^k (N_i - \bar{N})^2}{(k-1) \bar{N}}$$

where k = total of days and $\bar{N} = \sum_{i=1}^k N_i / k$ is the average daily count, is a classical test in this case. (We have divided the usual d by $(k-1)$

to suit our purposes). Under the hypothesis $\lambda_1 \equiv \lambda$, the expected value of d is unity. Smaller values indicate that less than sampling variation is present, and should not occur unless crowding is taking place (i.e., the hospital is reaching a maximum capacity for the type of disease considered). Values of d larger than unity occur if there is variation in the susceptibility to disease on different days. Exact critical values for d are given by Fisher [9] for small cell expectations, and $(k - 1) d$ is asymptotically a χ^2 variable with $(k - 1)$ degrees of freedom when all the N_i are large. When k is large but cell expectations are small, d is asymptotically normal with mean $(k - 1)$ and variance given by Morris [10]. Large values of d are indicative that the disease is affected by factors which are changing with time. These factors in the case of respiratory disease could include, for example, meteorological variables, air pollutants, epidemics, holidays, day of the week, and others.

When d is large, there are significant non-random daily variations. These variations may be of two general types: one consisting of slow seasonal or yearly trends, and the other consisting of rapid and more irregular fluctuations, such as day-to-day variations.

In order to estimate and then eliminate long trends, it is useful to test the λ_i for a fit to some simple curve such as a linear or quadratic fit over a season. (Data were never lumped over periods longer than three months; the seasons consisted of winter: December, January and February; spring: March, April and May, and so on). To test for short-term fluctuations when slower seasonal trends are present in the

data, a similar statistic to equation (1) was constructed, except that N was replaced by the trend line and the factor $\frac{k}{(k-1)}$ was adjusted, as necessary, to account for loss of degrees of freedom in fitting the trend line. For example, if we take the trend line to be a quadratic $\lambda_i = Ai^2 + Bi + C$, and A , B and C are estimated from the data, then the dispersion statistic d_2

$$d_2 = \frac{\sum_{i=1}^k (N_i - \lambda_i)^2}{(k - 3) \bar{N}}$$

is the appropriate statistic, and when all N_i are large, and $(k - 3) \cdot d_2$ will be asymptotically χ^2 with $(k - 3)$ df. Large values of d_2 will now indicate day-to-day variation in the N_i which cannot be explained by random variations about a slowly varying trend. The values for d_2 by season for nine seasons for asthma visits to the three hospitals in the two inner city areas of New York City for adults (over 13 years of age) and children (under 13 years of age) are given in Table 1. The greatest variation occurs consistently in the fall seasons.

Two Sample Cases

The geographical scale of possible environmental factors affecting respiratory illnesses can be determined by considering daily visits to different pairs of hospitals in the New York City area. From such comparisons, some clues can be obtained about the type of environmental factors involved: are these factors likely to prevail simultaneously throughout the City (we shall call them "global" factors) or do these factors occur locally in small geographic areas of the City affecting

only a limited population? If "global" factors are solely responsible for temporal variation in emergency room visits, then the underlying expected values at hospital A, λ_i^A , and hospital B, λ_i^B , on day i should have the same ratio on every day. Thus there would exist an (unknown) constant K such that

$$(2) \quad \lambda_i^A = K \lambda_i^B \text{ for all } i.$$

A suitable test for (2) is the χ^2 test for homogeneity on the $2 \times n$ table formed by the daily counts N_i^A and N_i^B at the two hospitals. When the counts N_i^A and N_i^B are all sufficiently large, this χ^2 statistic will have a χ^2 distribution with $(n - 1)$ df. When cell counts are small, but n is large, the asymptotic distribution of χ^2 is still normal with mean and variance given by Cuzick. [11]

Here large values of χ^2 indicate that factors affecting the temporal variation are local, whereas normal values indicate the disease is affected by a factor common to both areas serviced by the two hospitals, and thus acting on a larger spatial scale. Table 2 compares Harlem Hospital (in Manhattan) with Cumberland and Kings County Hospitals (in Brooklyn, in the general area of the Bedford-Stuyvesant district) for emergency room visits for asthma and upper respiratory complaints for monthly periods during the three years 1969-1971. Children under 13 are not considered except when comparing Cumberland and Harlem Hospital, because data for children were unavailable from Kings County. For 31-day months, the approximate 5% significance level for χ^2 is 43.8. For 30-day months, χ^2 is 42.6: Figure 1 shows histograms

of the monthly variations in three cases as follows: Figure 2B and 2C respectively, show the frequency distribution of monthly χ^2 values for Cumberland versus Kings County and Harlem versus Kings County Hospitals' adult emergency room visits for asthma. Figure 2D shows the frequency distribution of monthly χ^2 values for Kings County adult emergency room visits for upper respiratory complaints versus those at Harlem Hospital. Figure 2A shows the theoretical histogram under the null hypothesis of no difference between hospitals. It was constructed in the following way: It was assumed that (under the null hypothesis) the 36 monthly chi-square values form a sample from a chi-square distribution with 30 degrees of freedom. Since a chi-square distribution with 30 degrees of freedom can be reasonably approximated by the normal distribution, the normal was used to calculate the portion of the 36 chi-square values which are expected to fall in each of the intervals of the histogram. Figures 2B through 2D represent histograms which summarize the 36 monthly chi-square values for a given pair of hospitals, by diagnostic rubric and age group. We can see that the histogram representing upper respiratory complaints (Figure 2D) shows a much greater deviation from the theoretical histogram (Figure 2A) than do the histograms representing asthma visits to the emergency rooms (Figures 2B and 2C). Figure 2C which represents asthma visits in Harlem Hospital in Manhattan versus Kings County Hospital in Brooklyn shows greater deviation from the theoretical histogram (Figure 2A) than does the histogram representing asthma visits at the two hospitals situated in Brooklyn (Figure 2B).

Conclusion and Results

Table 3, which is a summary of Table 2, gives the average values of χ^2 obtained for each season separately for each year, and for each season averaged over the three years, for each pair of hospitals, for each of the three diagnostic categories considered. The following conclusions can be drawn:

1. The variation between the two hospitals located in Brooklyn are smaller than the variations between either of the Brooklyn Hospitals and the hospitals in Harlem for all diagnostic categories.
2. For asthma the variation between all pairs of hospitals is close to that predicted under the assumption that the factors affecting the visits to the emergency rooms are global in character.
3. The variations between hospitals for upper respiratory conditions are larger than for asthma and are clearly statistically significant.

There are a number of hypotheses that might be offered to account for the patterns observed in the temporo-spatial distributions of illness revealed by this analysis. For example, the fact that daily variations in upper respiratory illnesses about the trend line are non-random may be partly explained by the contagious aspect of some of the ailments. The concentration of non-randomness of asthma daily variations in fall seasons has been discussed in our previous publication. The fact that there is greater similarity between the Brooklyn Hospitals than between either Brooklyn Hospital and Harlem may reflect localized epidemics

either involving a contagious agent or another local environmental agent, or on the other hand, it may reflect different administrative practices among the hospitals. All of these are questions deserving further study.

The primary purpose of presenting these results was to provide an example of the possible application of a new methodology for the identification of possible environmental events that does not require a priori identification and measurement of environmental pollutants.

It should be clear that the method is of wide applicability.

References

1. Goldstein, I.F., Landovitz, L., and Block, G., "Air Pollution Patterns of New York City". Journ. of Air Pol. Cont. Assoc., 24:, No. 2, 148-152, February, 1974.
2. Goldstein, I.F., Landovitz, L. "Sulfur Dioxide: Harmful Pollutant or Air Quality Indicator?". Journ. of Air Pol. Cont. Assoc., 25:, No. 12, 1195, December, 1975.
3. Goldstein, I.F., Landovitz, L., "Analysis of Air Pollution Patterns in New York City: I. Can One Station Represent the Large Metropolitan Area?", Atmos. Environ., 11:, 47-52, January, 1977.
4. Goldstein, I.F., Landovitz, L., "Analysis of Air Pollution Patterns in New York City: II. Can One Aerometric Station Represent the Area Surrounding It?" Atmos. Environ. 11: 53-57, January, 1977.
5. Goldstein, I.F., "Scientific Base for Sulfur Oxide Air Pollution Standards". Proceedings of Sulfur Oxide Conference, John Carroll University Chair in Ecology, November, 1976.
6. Goldstein, I., "Discussion of Evidence for the Influence of Sulfur Oxides and Related Particulates on Mortality". "Evaluation of Research Strategies for Investigation of Health Effects of Air Pollution". Proceedings of the Symposium on Environmental Effects of Sulfur Oxides and Related Particulates. New York Academy of Medicine, March, 1978.
7. Goldstein, I., Rausch, L., "Time Series Analysis of Morbidity Data for Assessment of Acute Environmental Health Effects". Environ. Research, 1978.
8. Goldstein, I.F., Arthur, S., "Asthma Alley: A Space Clustering Study of Asthma in Brooklyn, New York City". Journ. of Asthma Research, 1978.
9. Cox, D.R. and Lewis, P.A.W., "The Statistical Analysis of Series of Events", p. 158, Methuen, London, 1966.
10. Morris, C., "Central Limit Theorems for Multinomial Sums", Annals of Statistics, 3: 346-355, 1975.
11. Cuzick, J. "Asymptotic Normality of χ^2 in $m \times n$ Tables with n large and Small Cell Expectations". Unpublished technical report, SIAM Institute for Mathematics and Society, 1977.

Table 1

DISPERSSION (d_2) OF DAILY NUMBER
OF VISITS FOR ASTHMA

Period	Number of Days	Cumberland		Harlem		Kings County
		<13 ¹	13+ ²	<13	13+	13+
9/69-11/69	91	2.46 ***	1.70 ***	1.24	1.82 ***	3.20 ***
12/69- 2/70	90	1.20	1.11	.91	1.51 ***	1.34 *
3/70- 5/70	92	1.34 *	1.06	1.17	1.30 *	1.08
6/70- 8/70	92	.89	1.22	1.26 *	1.04	1.30 *
9/70-11/70	91	1.36 **	1.09	1.49 ***	1.74 ***	1.90 ***
12/70- 2/71	90	1.48 ***	1.03	1.37 **	1.75 ***	1.49 ***
3/71- 5/71	92	1.56 ***	.92	1.21	1.67 ***	1.18
6/71- 8/71	92	1.08	1.06	1.43 **	1.42 **	1.50 ***
9/71-11/71	91	1.94 ***	1.97 ***	1.39 **	1.28 *	1.67 ***

1 - 13 years of age and under

2 - over 13 years of age

* Significant at 5% level

** Significant at 1% level

*** Significant at 0.1% level

Table 2

CHI-SQUARES FOR NUMBER OF EMERGENCY ROOM
VISITS IN PAIRS OF HOSPITALS BY MONTH

	CUMB. VS. K.C.		CUMB. VS. HARLEM				K.C. VS. HARLEM	
	Asthma	URI	Asthma		URI		Asthma	URI
	13+	13+	<13	13+	<13	13+	13+	13+
1969								
1	30.8	55.4		33.2		65.5	22.6	58.5
2	22.6	26.9		38.7		64.4	36.9	71.2
3	32.1	39.2		52.1		93.0	25.7	94.7
4	55.0	31.0		68.0		55.9	30.2	45.6
5	34.1	38.6		55.7		37.3	56.0	50.3
6	44.0	34.1		34.8		53.9	43.6	42.5
7	20.8	27.1		35.1		27.4	39.3	21.9
8	33.3	39.2	30.9	46.1	60.0	41.1	49.6	33.4
9	26.1	34.5	30.9	34.7	57.8	58.7	45.8	70.5
10	39.0	39.0	51.4	23.0	77.6	42.0	38.2	44.9
11	28.5	48.1	15.5	22.5	89.6	71.1	51.0	53.3
12	36.2	46.9	28.0	34.2	72.0	48.0	44.4	57.0
1970								
1	27.6	24.0	34.6	51.8	92.1	61.4	34.9	93.8
2	29.6	33.0	17.8	19.7	56.8	38.3	29.5	56.2
3	25.9	39.6	39.4	33.8	75.6	31.7	32.8	57.7
4	21.2	50.2	48.6	22.0	73.9	46.0	27.0	32.3
5	32.0	54.1	45.1	33.4	39.7	43.3	41.3	35.4
6	30.2	38.0	24.8	28.4	69.3	37.5	16.9	37.7
7	30.4	27.0	21.1	34.7	57.3	34.8	43.9	39.6
8	42.0	28.6	44.6	35.2	76.6	41.1	26.9	47.7
9	36.4	40.8	29.1	21.5	54.1	50.4	35.1	41.1
10	35.3	39.5	33.6	29.9	74.9	47.0	48.8	44.0
11	15.8	53.8	44.2	44.9	90.8	44.6	31.8	35.8
12	25.4	31.3	38.4	27.4	63.9	26.2	34.8	72.8
1971								
1	29.9	49.5	35.0	26.3	156.3	70.4	46.4	55.3
2	22.2	45.0	32.1	52.2	138.8	86.0	42.8	97.4
3	32.2	35.6	57.4	34.0	102.1	52.6	51.1	56.1
4	23.2	27.8	46.8	52.3	71.4	39.8	50.3	35.8
5	27.8	33.1	37.3	38.1	44.9	37.2	40.8	50.7
6	27.7	30.0	41.7	38.8	49.1	27.3	34.9	48.3
7	28.2	28.2	28.9	48.4	54.2	25.3	42.5	36.3
8	26.6	49.1	21.9	31.0	94.0	49.7	32.1	35.4
9	29.2	46.9	39.8	33.4	89.8	32.3	46.5	23.4
10	20.2	40.2	40.0	34.8	94.6	42.2	29.2	36.8
11	36.6	29.5	58.1	30.0	79.9	63.2	28.5	99.9
12	22.7	41.6	45.2	28.3	81.6	45.8	41.8	69.6

1. Cumberland Hospital (Brooklyn) 2. Kings County Hospital (Brooklyn)
3. Harlem Hospital (Manhattan)

95% Significance levels:

43.8 for 31 day months; 42.6 for 30 day months; 40.1 for 28 day months

Table 3

CHI-SQUARES FOR NUMBER OF EMERGENCY ROOM VISITS
FOR PAIRS OF HOSPITALS AVERAGES OVER SEASONS

	<u>ASTHMA</u>				<u>UPPER RESPIRATORY¹</u>			
	1969	1970	1971	Average	1969	1970	1971	Average
				<u>CUMB.³ VS. K.C.⁴</u>				
Winter		31.1	25.8	28.5		34.6	41.9	38.3
Spring	40.4	26.3	27.7	31.5	36.3	48.0	32.2	38.8
Summer	32.7	34.2	27.5	31.5	33.5	31.2	35.8	33.5
Fall	31.2	29.2	28.9	29.8	40.5	44.7	38.9	41.4
				<u>CUMB. VS. HARLEM</u>				
Winter		35.2	35.3	35.3		49.2	60.9	55.0
Spring	58.6	29.8	41.5	43.3	62.0	40.3	43.2	48.5
Summer	38.7	32.8	39.4	37.1	40.8	37.8	34.1	37.6
Fall	26.7	32.1	32.8	30.5	57.2	47.3	46.0	50.2
				<u>K.C. VS. HARLEM</u>				
Winter		36.3	41.3	38.8		69.1	75.2	72.1
Spring	37.3	33.7	47.4	39.5	63.5	41.8	47.5	51.0
Summer	44.2	29.2	36.5	36.6	32.6	41.7	40.0	38.1
Fall	45.0	38.6	34.7	39.4	56.2	40.3	53.4	50.0
				<u>CUMB. VS. HARLEM (under 13 years of age)</u>				
Winter		26.8	35.2	31.0		73.6	119.7	96.6
Spring		44.4	47.2	45.8		63.1	72.8	68.0
Summer		30.1	30.8	30.5		67.7	65.8	66.7
Fall	32.6	35.6	46.0	38.1	75.0	73.3	88.1	78.8

1. Visits to emergency rooms coded under upper respiratory diagnoses

2. Average for season for all years. 3. Cumberland Hospital

4. Kings County Hospital

95% Significance levels:

Yearly for each season 37.5; Average for season for all years 34.1.

Figure 1

Fig. Aerometric network -- New York City.

Figure 2

HISTOGRAMS SUMMARIZING 36 MONTHLY CHI-SQUARE
VALUES FOR PAIRS OF HOSPITALS

- 2A. Theoretical histogram under the null hypothesis of no difference between the hospitals.
- 2B. Adult emergency room visits for Asthma at Cumberland versus Kings County Hospitals in Brooklyn.
- 2C. Adult emergency room visits for Asthma at Harlem Hospital in Manhattan versus Kings County Hospital in Brooklyn.
- 2D. Adult emergency room visits for Upper Respiratory complaints at Kings County versus Harlem Hospital.