REPORT NO 2/1984

NATIONAL INSTITUTE OF
ENVIRONMENTAL MEDICINE

STRATEGIES FOR PATTERN RECOGNITION OF AIR SAMPLES FROM NORMAL AND SICK BUILDINGS

JOHN C. BAIRD, BIRGITTA BERGLUND, ULF BERGLUND,
BERNARD DEVINE AND HÉLÈNE NICANDER-BREDBERG

DEPARTMENT OF PSYCHOLOGY, UNIVERSITY OF STOCKHOLM,
DEPARTMENT OF ARCHITECTURE, THE ROYAL INSTITUTE OF TECHNOLOGY
AND DEPARTMENT OF HYGIENE, NATIONAL INSTITUTE OF ENVIRONMENTAL
MEDICINE, STOCKHOLM

National Institute of Environmental Medicine
S-104 01 STOCKHOLM

Report No. 2/1984

STRATEGIES FOR PATTERN RECOGNITION OF AIR SAMPLES FROM NORMAL AND SICK BUILDINGS*

Baird, J.C., Berglund, B., Berglund, U.,
Devine, B., and Nicander-Bredberg, H.

The Swedish building code of 1975 emphasizes energy conservation and encourages the construction of tightly insulated structures with adequate ventilation systems. Some of the new buildings constructed along these lines have been labeled "sick", because people working in them report an unusual number of health problems - e.g., eye irritation, skin rashes, and colds. One possible indicator of whether a building is "normal" or "sick" may be looked for in the pattern of chemicals present in the air. This article outlines four strategies of analysis designed to find and recognize chemical patterns associated with each type of building and consequently to identify the physical materials responsible for emitting the chemicals involved in "sick building" environments. The approaches combine statistical, clustering, and scaling methods in ways which, as far as we know, have never been tried on this or related problems of pattern recognition and identification.

INTRODUCTION AND PROBLEM

Over recent years the occupants of certain buildings in Sweden have contracted health problems and reported symptoms similar to those caused by formaldehyde exposure, although the concentrations of formaldehyde in the indoor air were far below the accepted reaction thresholds (e.g., Andersen, Lundqvist & Mölhave, 1975). These buildings have been classified as an "irritating" type of sick building, and as such can be distinguished from other types of problem buildings, such as those contaminated with radon, molds or contagious agents. The former buildings include preschools, offices, health centers, private dwellings, and

---

*This research was supported by the Swedish Council for Building Research and the Swedish Council for Research in the Humanities and Social Sciences. The work was accomplished while Professor John C. Baird from the Departments of Psychology and Mathematics/Social Sciences at Dartmouth College in the USA was a visiting professor at the Department of Psychology, University of Stockholm and the Department of Hygiene of the National Institute of Environmental Medicine.

homes for the elderly. The majority of these "irritating" sick buildings in Sweden were constructed according to the Swedish building code of 1975 that placed a premium on saving energy. The first known cases of sick buildings appeared in 1977, but the magnitude of the problem has grown since then. For example, about 100 of the 600 preschools built most recently in greater Stockholm are affected.

The occupants of these sick buildings complain of poor air quality and show subtle medical symptoms that may be related to the chemical composition of the indoor air (Berglund, Berglund & Lindvall, 1984). The symptoms reported vary widely, but some common features are evident such as: (a) irritation of the eyes, nose and throat, (b) dry sensations in the mucosa and skin, (c) erythema, (d) mental fatigue, and (e) the perception of weak but persistent odors. It should be noted that the symptoms refer primarily to sensory reactions or perception of the environmental air rather than to some known medical illness. In addition to these common features, there are also cases of hoarseness of the throat, allergies, skin rashes, headaches, colds, excessive thirst, and loss of hair.

Most of the information about the sick buildings comes from occupational safety and health control agencies, and so far only a few epidemiological studies have been done. All attempts to isolate a specific chemical or infectious agent have failed. Moreover, an analysis of the problem in terms of possible psychogenic origins has not clarified the issue (for an overview see Berglund & Lindvall, 1983). Part of the difficulty in tracing chemical involvement is the large number of chemicals that must be examined. Each year in Sweden hundreds of new building materials are introduced, and 5000 - 6000 products alter their chemical content to some degree (Tell, 1983; U.S. National Research Council, 1981).

At first sick buildings were believed to contain an unusually high concentration of formaldehyde since the symptoms reported resemble those caused by formaldehyde - the other most common causes for poor indoor air quality having already been ruled out; namely, high room temperature or low humidity. Measurements of the concentrations of formaldehyde showed this was not the case so an alternative suggestion was that this might be an entirely new type of psychogenic illness.

However, since the chemical analyses done to date of indoor air have not produced any satisfactory explanation for the sick building syndrome, we feel it has become necessary to adopt a completely different approach to the problem. Instead of searching for a single target or small group of chemicals, we propose to seek an explanation in terms of more complex patterns of pollutants that might exist in the indoor air. The central idea is that certain combinations of pollutants could create interactions that can be the cause of the symptoms of ill-being.

Because most of the symptoms are sensory in nature, it seems reasonable to assume that human sensory systems (the sense of smell and/or the skin senses) perform a sort of pattern recognition analysis creating a multisensory image of the air. Consequently, one of the main aims in our research is to discover if there exist special combinations of chemicals that act as stimuli for human multisensory evaluation of air quality.

Types of chemical patterns

Different types of chemical and perceptual patterns can be obtained from air samples. They may be in the form of gas chromatograms (GC), mass spectra (MS), or percept-olfactograms (POG). They may also consist of patterns constructed from various combinations of statistical measures taken on the sample profiles. The procedure for chemically analyzing the air samples for obtaining the GC patterns is as follows.

An air sample (15 l) is adsorbed on a porous polymer filter. The sample is then desorbed into a stream of helium. The gas passes through a capillary coil which is chilled by liquid nitrogen. This injection system "trap" causes all the components with freezing points above the temperature of liquid nitrogen (-196 deg. C.) to solidify, whereas the helium gas passes straight through. In this way the sample is concentrated into a small volume, which is necessary to enable it to be injected into the gas chromatograph. Once inside the gas chromatograph the sample is separated in a capillary (stainless steel or fused silica) column. This method is described in detail by Johansson (1978). Part of the eluate (14%) is passed on to the flame ionization detector (FID), resulting in the gas chromatogram, and the remaining part (86%) is passed to human observers who make odor evaluations every 15 seconds during the 70-minute chemical analysis of the air sample (see Berglund, Berglund, Lindvall & Nicander-Bredberg, 1982). Percept-olfactograms are based on the human evaluations. We have found that the FID detects a larger number of chemical components than is evident in perceptual reports. It is also possible to obtain reports of eye irritation after exposing the eye to the eluate so that percept-irritograms can be constructed.

In addition to determining the chemical content of the air sample with the FID, we have also identified the substances with the aid of a mass spectrometer. Using the techniques available for the mass spectrometer one obtains gas chromatograms as well as mass spectra, the GC's from the mass spectrometer being obtained by total ionization detection. Using these GC-MS methods we can detect chemical substances within the range of 40 to 200 amu (atomic mass units) which have boiling points between 40 and 250 degrees C. Such analyses produce sample profiles that can be treated as individual patterns or as composites of smaller subpatterns.

ANALYTICAL STRATEGIES

The ultimate aim of this research is to identify the chemicals associated with the health problems reported in so called "sick" buildings. The plan to accomplish this compares the patterns of chemicals found in "sick" buildings with the corresponding ones obtained from so called "normal" buildings. In this document we describe four possible strategies for analyzing the chemical patterns. All four approaches are predicated on the procedural details described above and on some broad assumptions concerning the buildings, the chemicals, and the data collection methods. The following general conditions apply in all cases. Air of a number of sick and normal buildings have been sampled over periods of time (2 - 4

weeks). The samples were collected from specific locations including the outdoor air surrounding the building, the air in the ducts of the ventilation system, the air in the rooms of the building itself, and the air leaving the building from the main exhaust. The measurements carried out on the samples consist of the gas chromatogram, the mass spectrum, and human judgments of the magnitude of odor strength (percept-olfactogram). Parallel samples, secured at the same point in time on each occasion, were obtained in order to check the reliability of the measurement procedures. Such reliability can be assessed by correlation coefficients in the usual manner. The analyses proposed below can be done on any type of the sample profiles, and consequently, we will not make a point of mentioning them all every time a new analysis is introduced. The term "spectrogram" or "chromatogram" will be applied in a generic sense to refer to any type of spectral measurement.

STAGE 1

The goal in Stage 1 is to devise a prototypical pattern for each type of building and to reduce the number of chemicals that need be considered in distinguishing between "sick" and "normal" buildings. These prototypes can then be used to classify new samples based on measurements of their critical aspects, and can be used to assist in the identification of the environmental substances that emit such chemicals into the air. This approach is also central to all the following stages proposed here, so we shall describe it in some detail. Stage 1 can be partitioned into four distinct phases: (1) the separation of building and chemical types based on similarity, (2) the identification of the chemicals necessary to distinguish between buildings, (3) the definition of prototypes, and (4) the identification of the physical materials in the environment responsible for emitting the critical chemicals.

Phase 1 - Classification of buildings and chemicals

Two proximity matrices are computed; one for comparisons between building samples and the other for comparisons between chemicals. It is possible to employ a variety of measures from the individual spectrograms to construct such matrices. We propose beginning with the simplest of all similarity measures. First a list is made of all the different chemicals that are found during the chemical analysis of the population samples. Then the presence or absence of these specific chemicals within two spectrograms are taken as an indication of the similarity (proximity) between the spectrograms. A simple matching procedure can be employed, where similarity is computed as the number of chemicals present or absent in both samples, divided by the total number of comparisons for all the samples. Subsequent measures can include correlation coefficients between concentrations within two spectrograms and the average difference between standard scores (z scores) obtained for each component of a spectrogram.

The proximity matrices may serve as input for three types of analysis. Hierarchical cluster analysis (Baird & Noma, 1978; Johnson, 1967) is a good candidate to start with, the results of which offer a readily interpretable visual picture of the nested groupings of items (either building samples or chemicals). Hypothetical examples of these dendro-

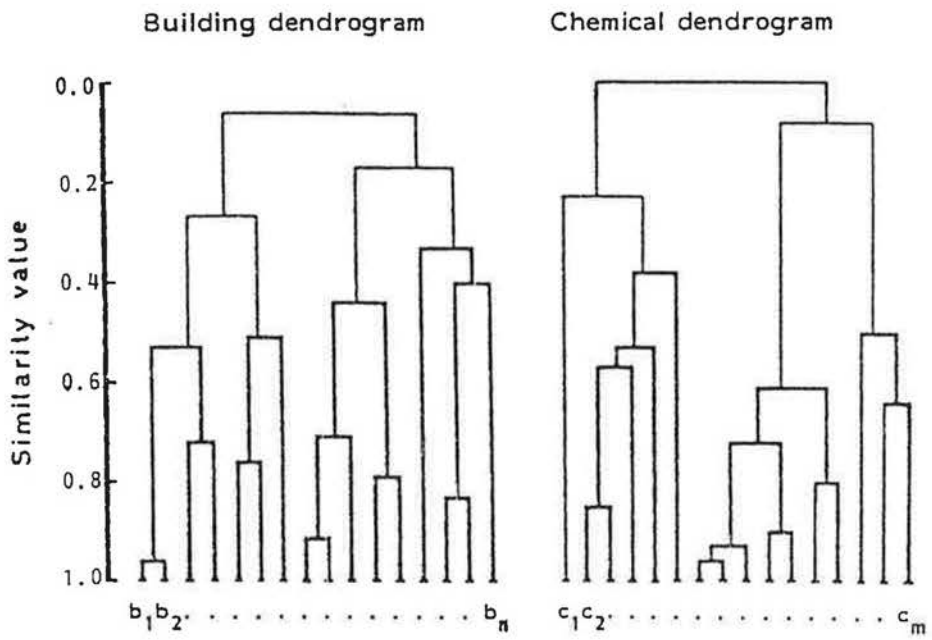

Fig. 1. Hypothetical results of hierarchical cluster analyses for buildings and chemicals. Different subgroups of chemicals can be removed in successive runs of the analysis on buildings.

gram results are shown in Fig. 1 for a limited number of buildings and chemicals. If cluster analysis does not distinguish satisfactorialy between groupings, then factor analysis and discriminant analysis (Gnandesikan, 1977; Harman, 1967; Jöreskog, Klovan, & Reyment, 1976) can be tried.

It is absolutely essential that a sound technique be found to separate sick and normal buildings. After this an iterative procedure can be started to determine which specific chemicals are important in discriminating between building types. This may be done as follows. Based on the chemical dendrogram (e.g., Fig. 1), remove a subcluster of items from the data base and then repeat the cluster analysis on building samples. If the sick and normal distinction is less clear in the latter dendrogram and the distinction evaporates between building types, then the deleted chemicals could be critical. Alternative approaches can be tried for determining the optimum sequence for removing chemicals. Our initial thoughts are that large clusters of items should be dropped first and then one should gradually move down the hierarchy to smaller and smaller clusters. This iteration procedure can continue until all the critical chemicals have been found, following which only this group of chemicals needs to be considered.

Phase 2 - Identifying chemicals and their groupings

At this juncture chemical analyses will be undertaken to identify specific critical chemicals and to find out which building materials emit them.

In order to facilitate this identification, multidimensional scaling (nonmetric) can be employed to supplement the dendrogram (Baird &

Noma, 1978; Coxon, 1982; Kruskal, 1964; Shepard, 1962). The rationale here is that specific "chemical dimensions" may be identified and interpreted, thus helping us recognize characteristic groups of the critical chemicals. The analyses can be carried out on the same proximity matrices as were used for the cluster analyses.

Phase 3 - Defining air prototypes

The next step is to look for a pattern that is typical for each of the building types (sick and normal). Two approaches may be employed. In the first approach the average concentrations are computed across all samples of a given building type to obtain a centroid pattern which can serve as the prototype. This is a standard method in statistical pattern recognition (Andrews, 1972; Varmuza, 1980). The second approach investigates the utility of defining a prototype in quite another manner. In this instance, all the samples from one type of building are compared with each other, and those components found to have the same concentration (within a designated confidence interval) for two or more samples become part of the prototype, regardless of the concentration levels present in other samples. The emerging prototype will probably be different from the one obtained using the first approach, which directly averages over all the samples. The motivation for the second approach is that temporary chemical "noise" in the air on a particular occasion may have a large impact on the results obtained from a specific sample in a series, the difference arising from the addition and removal of objects in the environment, either inside or outside the building. If under transient circumstances a component is found to have the same concentration level on at least two different occasions, then it seems reasonable to assume that this component is a stable, invariant part of the air mixture. Of course, one can enforce a stricter criterion - e.g., that the component concentration must match in three or more samples.

Once reliable prototypes are developed, it may be possible to weight the components and build a classifier system to assign new samples to either "sick" or "normal" categories. This latter phase can then unfold according to the usual procedures employed in chemical pattern recognition (Varmuza, 1980).

More recently, alternative statistical procedures to the ones described above have been applied to similar problems of chemical pattern recognition (e.g., Wold & Sjöström, 1977). These may also be tried in the present case.

Phase 4 - Identification of physical objects and substances

Armed with a set of key components and defined prototypes (chemical patterns) an attempt can be made to isolate the materials and objects in the environment that are consistent with the similarities and differences between sick and normal prototypes and their constituent chemicals.

STAGE 2

The aim in Stage 2 is to pinpoint specific chemicals that can be used to separate sick from normal buildings. Representations can be

shown visually to allow us to get a general picture of the chemical relationships, and to see which chemicals might be the relevant ones for securing a separation between the buildings.

Phase 1 - Obtaining visual displays

The concentration of the components from each chromatogram can be compared in visual diagrams. One way of accomplishing this is by plotting the rank order of the components from pairs of samples against each other. An illustrative example of such a plot is shown in Fig. 2. Various subsets of the data can be depicted; for example, only components that are present in both samples, or only components reaching a certain level of concentration. Displaying the differences between ranks, instead of the ranks themselves, may provide a still better means for comparing samples. In all instances it is necessary to calculate correlation coefficients for the entire spectrogram and appropriate measures of deviation for each component chemical (e.g., median difference in rank between each chemical component in the sick and normal buildings, average squared deviations in concentration for each component). These statistics are then tabled for visual reference. It is also possible to utilize the dendro-

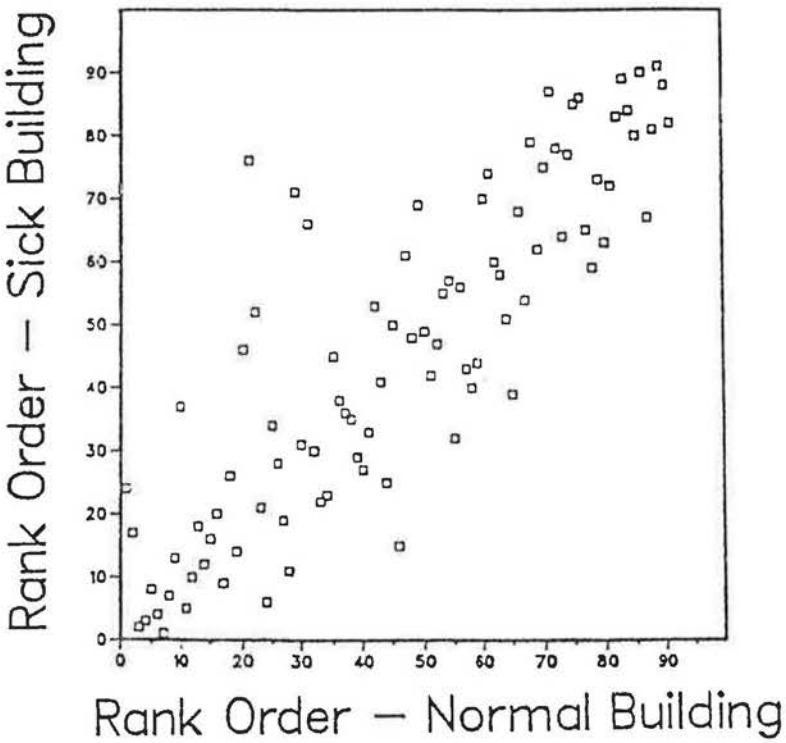

Fig. 2. Comparison of rank order concentration for individual mass numbers (summed mass spectrum) in two types of buildings (normal, sick). Data points represent real outcomes.

grams from Stage 1 to obtain plots of the final rank orders of the chemical components to compare different types of buildings. This might prove to be more meaningful than comparing the components with regard to their concentration, since the behavior of the chemicals across all the building samples is considered.

Phase 2 - Separating buildings and chemicals

The importance of each chemical can be evaluated on the basis of the outcome of Phase 1. In particular, chemicals that turn up in both sick and normal buildings should be tentatively removed, and the chemicals showing the largest discrepancies between the two building types can then serve as input to the analyses described in Stage 1. It must be remembered here that there is the possibility of combination effects; that is, that one chemical might have an effect only in the presence of another, or several others.

STAGE 3

The purpose of Stage 3 is to consider a variety of measures simulataneously, if it turns out that the patterns arising from the analyses of the spectrum are of such complexity that simple methods do not give any indication of which chemicals might be responsible for sick buildings.

Phase 1 - Statistics

Various statistics can be computed for each gas chromatogram including the total and mean concentration across all the components, the standard deviation of the concentration and the relative error (standard deviation divided by the mean), the number of nonzero components, and perhaps, some higher moments (e.g., skewness and kurtosis of the distributions).

Proximity matrices, composed of similarity measures, can be created based on each statistic individually or in combination with others. For example, if each set of N statistical measures were divided into a higher and lower category, a single chromatogram would be symbolized as an N-bit number, where "1" designated presence in the upper half of the distribution and "0" designated presence in the lower half. For instance, a sample might be represented by strings such as 10110 or 00111. The similarity measure between chromatograms could be computed by treating each digit as a discrete feature, and overall similarity assessed in the same manner as discussed in Stage 1.

A spatial example of how one might organize a set of spectrograms based on two statistics (mean and standard deviation of peak concentration) is given in Fig. 3.

Phase 2 - Reapply the statistical methods used in Stage 1-Phase 1 to the new proximity matrices.

STAGE 4

The aim in Stage 4 is to study transfer functions for each type of building. At this stage a single building and its immediate environment can be treated in its totality as a system where there is an exchange of air between parts of the system. There are four sorts of air samples for each building: outdoor, indoor at the ventilation inlet,

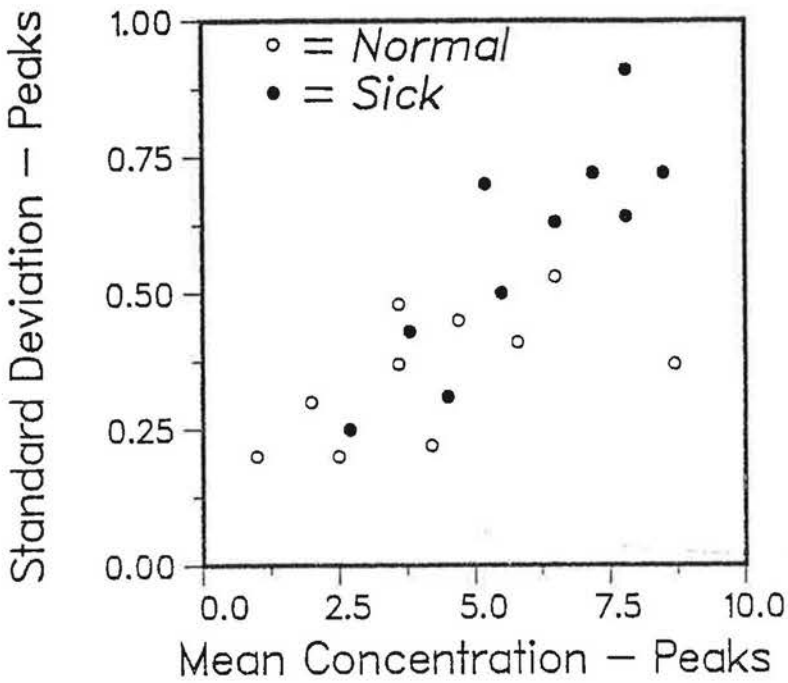

Fig. 3. Hypothetical spectrograms of the air in 10 normal and 10 sick buildings, as represented by the mean and standard deviation of their respective peak concentrations.

indoor within the room, and at the main exhaust. At each of the interfaces between indoor and outdoor air, there is a filter and/or ventilation mechanism that changes the nature of the air mixture in characteristic ways. It is possible that the basic components of the air associated with sick and normal buildings are essentially the same, but imperfections and/or differences at the interfaces alter the chemical milieu so that health problems arise. If this is indeed true, methods should be sought to find the transfer function that best describes the relation between the air samples on both sides of each interface.

Phase 1 - Determining transfer functions

The search for transfer functions can be started by employing each of the statistics (or selected subgroups) computed in Stage 3. In the first instance single values of a statistic can be used (e.g., mean concentration) for each of the samples obtained for a category of interest (e.g., normal buildings). The value obtained for one side of the interface can be regressed against the comparable statistic found on the other side. Curve-fitting procedures may be employed to determine the best-fitting functions for each case. Both the nature of each resultant function and the parameter values can be compared between normal and sick buildings. If differences are found this could indicate that at least part of the problem could lie in the nature of the physical interfaces. A hypothetical example of a transfer function is shown in Fig. 4.

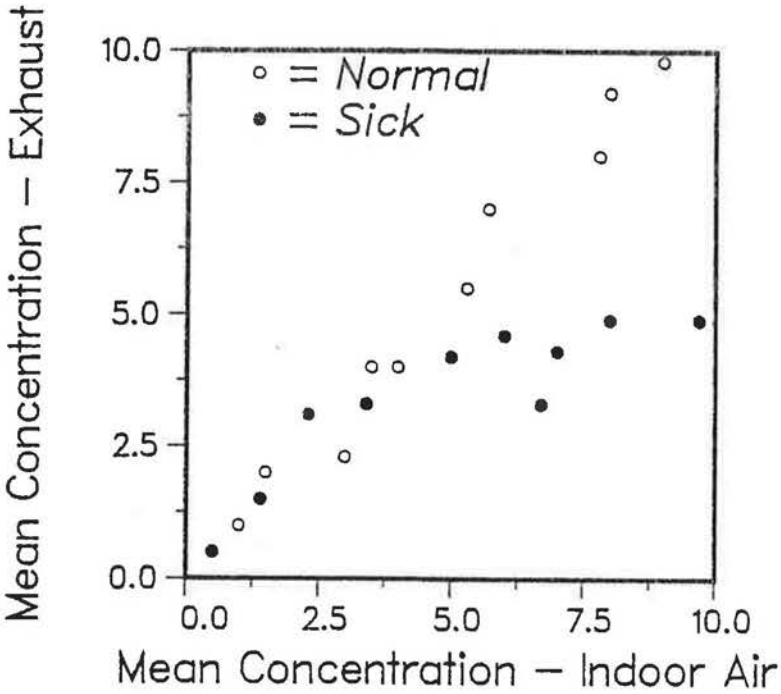

Fig. 4. Hypothetical data from 10 normal and 10 sick buildings, as represented by the relation between the mean concentration of their indoor and exhaust air.

PRELIMINARY ANALYSIS

The first step in this research program will be to produce a typical, but greatly compressed, artificial data set to simulate the expected empirical samples to be analyzed. In this way we will develop and run all the necessary computer programs on a manageable data set, and should thereby achieve a better feeling about the reasonableness of the proposed techniques; in particular, their sensitivity to changes in the number of variables, mean, variance, and so forth. If none of the foregoing approaches is successful on this reduced scale, it will be time to rethink the problem in the light of any insights gained from the analyses conducted up to this point.

FUTURE CONSEQUENCES

The chief purposes of this research are to diagnose "sick" and "healthy" buildings and to find which materials are emitting substances that give rise to unhealthy building environments. Because it appears unlikely that a single chemical is responsible for the health problems reported to date, we anticipate a solution in terms of a "symptom pattern" of different chemicals. Consequently a building could be evaluated by a set of scores on a battery of tests, and it would be the profile across the battery that would determine its eventual designation as "sick" or "normal". This approach follows the conventional

wisdom of psychological testing - one is seldom able to distinguish among individuals based solely on one measure secured by a single test.

If an identification of the materials causing the problems can be accomplished, it should be possible to suggest precautionary measures by referring to the air prototypes and comparing them against the pattern of chemicals emitted by the raw materials used in new constructions.

Finally, the goal of the present type of research must be the establishment of standards to ensure the safety and health of the tenants and occupants of all building types.

REFERENCES

Andersen, I., Lundqvist, G.R., and Mölhave, L. Indoor air pollution due to chipboard used as a construction material. Atmospheric Environment, 1975, 9, 1121-1127.

Andrews, H.C. Introduction to Mathematical Techniques in Pattern Recognition. New York: Wiley-Interscience, 1972.

Baird, J.C., and Noma, E. Fundamentals of Scaling and Psychophysics. New York: Wiley-Interscience, 1978.

Berglund, B., Berglund, U., and Lindvall, T. Characterization of indoor air quality and "sick buildings". ASHRAE Transactions, 1984, 90, part 1, 1045-1055.

Berglund, B., Berglund, U., Lindvall, T., and Nicander-Bredberg, H. Olfactory and chemical characterization of indoor air. Towards a psychophysical model for air quality. Environment International, 1982, 8, 327-332.

Berglund, B., and Lindvall, T. Sensory reactions to "sick buildings". In B. Berglund & C. Levy-Leboyer (Eds.), Applications of Environmental Psychology: Recent Research on Environmental Hazards and Unfavorable Environments. London: Sage Publications, 1983. (In press)

Coxon, A.P.M. The User's Guide to Multidimensional Scaling. London: Heinemann Educational Books, 1982.

Gnanadesikan, R. Methods for Statistical Data Analysis of Multivariate Observations. New York: John Wiley & Sons, 1977.

Harman, H.H. Modern Factor Analysis. London: University of Chicago Press, 1967 (2nd ed).

Johansson, I. Determination of organic compounds in indoor air with potential reference to air quality. Atmospheric Environment, 1978, 12, 1371-1377.

Johnson, S.C. Hierarchical clustering schemes. Psychometrika, 1967, 32. 241-254.

Jöreskog, K.G., Klovan, J.E., and Reyment, R.A. Geological Factor Analysis. New York: Elsevier, 1976.

Kruskal, J.B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. I, II. Psychometrika, 1964, 29, 1-27.

Shepard, R.N. The analysis of proximities: Multidimensional scaling with an unknown distance function. I and II. Psychometrika, 1962, 27, 125-140, 219-246.

Tell, W. Hälsofarliga byggnadsmaterial. VVS, Special No. 2, 1983, 43-46.

U.S. National Research Council. Indoor Pollutants. Washington, D.C.: National Academy Press, 1981.

Varmuza, K. Pattern Recognition in Chemistry. New York: Springer-Verlag, 1980.

Wold, S., and Sjöström, M. SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy. In B. R. Kowalski (Ed.) Chemometrics: Theory and Application. American Chemical Society, (ACS) Symposium Series, No. 52, 1977.