



ELSEVIER

Environmental Modelling & Software 16 (2001) 525–532

Environmental
Modelling & Software

www.elsevier.com/locate/envsoft

The environment and the quality of life in the United States over time

J.G. Hirschberg ^{a,*}, E. Maasoumi ^{1,b}, D.J. Slottje ^b

^a *Department of Economics, University of Melbourne, Parkville, 3052, Australia*

^b *Department of Economics, Southern Methodist University, Dallas, TX 75275, USA*

The environment and the quality of life in the United States over time

J.G. Hirschberg ^{a,*}, E. Maasoumi ^{1,b}, D.J. Slottje ^b

^a Department of Economics, University of Melbourne, Parkville, 3052, Australia

^b Department of Economics, Southern Methodist University, Dallas, TX 75275, USA

Abstract

This paper considers the similarity between two measures of air pollution/quality control, on the one hand, and widely used indicators of life quality and welfare in the US, on the other. We use statistical cluster analysis based on information measures of similarity and distance. The results indicate the clusters of the US macroeconomic and quality of life indicators from 1915 to 1987, which contain the air quality measures. This can help in deciding on distinct dimensions to be considered in multidimensional studies of welfare and quality of life. © 2001 Elsevier Science Ltd. All rights reserved.

JEL classification: C82; I31; C14

Keywords: Air quality; Time series; Information measures; Economic welfare; Entropy

1. Introduction

Quantitative analysis of quality of life across countries, and the construction of summary indices for such analyses has been of interest for some time, cf. Slottje et al. (1991) for a discussion. Most early work focused on largely single dimensional analysis based on such indicators as per capita GNP, the literacy rate, and mortality rates. In the 1980s, Maasoumi (1986) and others called for a multidimensional quantitative study of welfare and quality of life. The argument is that welfare is made up of several distinct dimensions, which cannot all be monetized, and heterogeneity complications are best accommodated in multidimensional analysis. See Maasoumi (1998) for a recent survey.

Hirschberg et al. (1991, 1998) introduced cluster analysis in this area as a means of identifying similar indicators, and collecting them into distinct clusters which could represent the dimensions worthy of distinct treatment in multidimensional frameworks. They analysed clusters of welfare attributes across countries, utilizing their cross section distribution and clusters of the

same over time in a given country, utilizing the time series distribution of many quality of life attributes and economic indicators.

The innovation in this paper is to consider the role of air quality indicators in the context of the cluster analysis outlined in Hirschberg et al. (1998). We identify the distances and similarities between the two air quality indicators and 15 other welfare attributes. This should help analysts to decide if other closely related attributes are representing air quality influences, thereby avoiding double counting of the same quality dimensions, or deciding to replace other indicators with measures of air quality without loss of “statistical information”.

Concurrently, another objective of this paper is to introduce the reader to a technique that allows the comparison of various attributes that impact the quality of life in the United States over time in a meaningful way. Specifically, cluster analysis techniques and kernel estimation are used in this paper. The methods used here were developed in our earlier paper (Hirschberg et al., 1998). In Section 2 of the paper, we discuss the attributes to be analyzed in the study and discuss the role of the environment in the quality of life. Section 3 describes our cluster methodology and the software used here and, Section 4 presents the empirical results. Section 5 concludes the study.

* Corresponding author. Tel.: +61-3-9344-5273.

E-mail address: j.hirschberg@unimelb.edu.au (J.G. Hirschberg).

¹ Tel.: +61-3-8344-5273.

2. The environment and other attributes of economic well-being

It is well known that the quality of the air in a locale influences the health of the population and ultimately effects other dimensions of that population's welfare and its economy. As a simple example, in cities where pollution levels rise significantly in the summer, worker absenteeism rates rise commensurately and productivity is adversely impacted. Other dimensions of the economy are influenced on "high pollution days" as well. For example, when outdoor leisure activity is restricted this may have serious consequences for the service sector of the economy (see Bresnahan et al., 1997 for an analysis of how activity is restricted by heavy air pollution). In this paper, we introduce two measures of environmental quality or air quality as quality of life factors.

A feature of these indices is the fact that these types of pollution are created by some of the very activities that define economic development. The two factors under investigation here are sulfur dioxide (SO₂) and nitrogen oxide (NO_x). They are both produced as by-products of fuel consumption as in case of the generation of electricity. Vehicle engines also produce a large proportion of NO_x. SO₂ is primarily produced when high sulfur coal is burned which is usually in large-scale industrial processes and power generation. These series are generated from production figures that are multiplied by an "emission factor" which are reported in detail by the US EPA (1998b). Thus, the ratio of these emissions to the population is an indication of pollution control. These two pollutants are examined here because the projected series are so detailed which is necessary for the analysis applied in this case.

The attributes of quality of life used in this paper are basic measures of wellbeing; they include the factors described in Table 1 below.² Note that these factors are defined in such a way that increases in these variables would be interpreted as an increase in the level of welfare. Thus, the measures of pollution have been defined as the inverse of tons per capita. These series were made available from the US EPA (1998a).³ When clustering disparately measured variables, it is necessary to define the variables in such a way that the relative variations in the variables are comparable.

The plot of each time series for each attribute is given in Fig. 1. All of these attributes have been rescaled to have a mean of zero and a standard deviation of one. Note from this figure that the control of NO_x (L16)

² See Hirschberg et al. (1998) for a more detailed discussion of these factors.

³ The methods used in the construction of this data are given in "National Air Pollutant Emission Trends Report, 1900–1996". www.epa.gov/oar/emtrnd. The 1947–1985 data were obtained directly from a Mr Tom McMullen of the EPA.

Table 1
The factors used to characterize the quality of life and air pollution

Factor	Description
L1	Per capita GNP
L2	1/infant mortality
L3	Male life expectancy
L4	Female life expectancy
L5	Employment rate
L6	Per capita Disp income
L7	Physicians/100,000
L8	Hyw miles per capita
L9	Homes with phones (%)
L10	Homes with radio (%)
L11	100,000/Homicide
L12	Percentage age 5–17 in school
L13	Newspapers/capita
L14	GNP rate of growth
L15	Percentage GNP not defense spending
Air pollution measures	
L16	Population/ton of NO _x
L17	Population/ton of SO ₂

seems to be in decline while the control of SO₂ (L17) is getting better. Section 3 defines the distance metric used in the clustering of these series.

3. Cluster analysis

The concept of clustering time series has been the subject of a number of recent papers (i.e. Piccolo, 1990; Maharaj, 1995). These authors propose grouping time series by the "similarity" of the parameters of ARIMA models fit to the time series. A difficulty with these studies is that they assume that a parameterised model explains the series under review entirely. Another problem with these methods in the present case is that some of our series are integrated and some are not. But comparisons across ARIMA models of different orders of integration involve a comparison of dissimilar parameters. In addition, the method proposed by Maharaj is not a metric since it violates the triangle inequality (see Kaufman and Rousseeuw (1990, p. 13)) when it uses the estimated covariances of the parameters as well as the parameter estimates. The violation of the triangle inequality means that the distance between series A and C can be greater than the distance from A to B plus the distance from B to C. Thus, one is measuring divergence not "distance", and cannot use geometric analogies to interpret the results. In this paper we follow the methodology proposed in Hirschberg et al. (1998) in which an information theoretic entropy metric is used to establish the distance between time series based on the similarities of the estimated conditional *distribution* of these series. In our approach, we first determine the systematic part of each series by fitting them to parametric time series

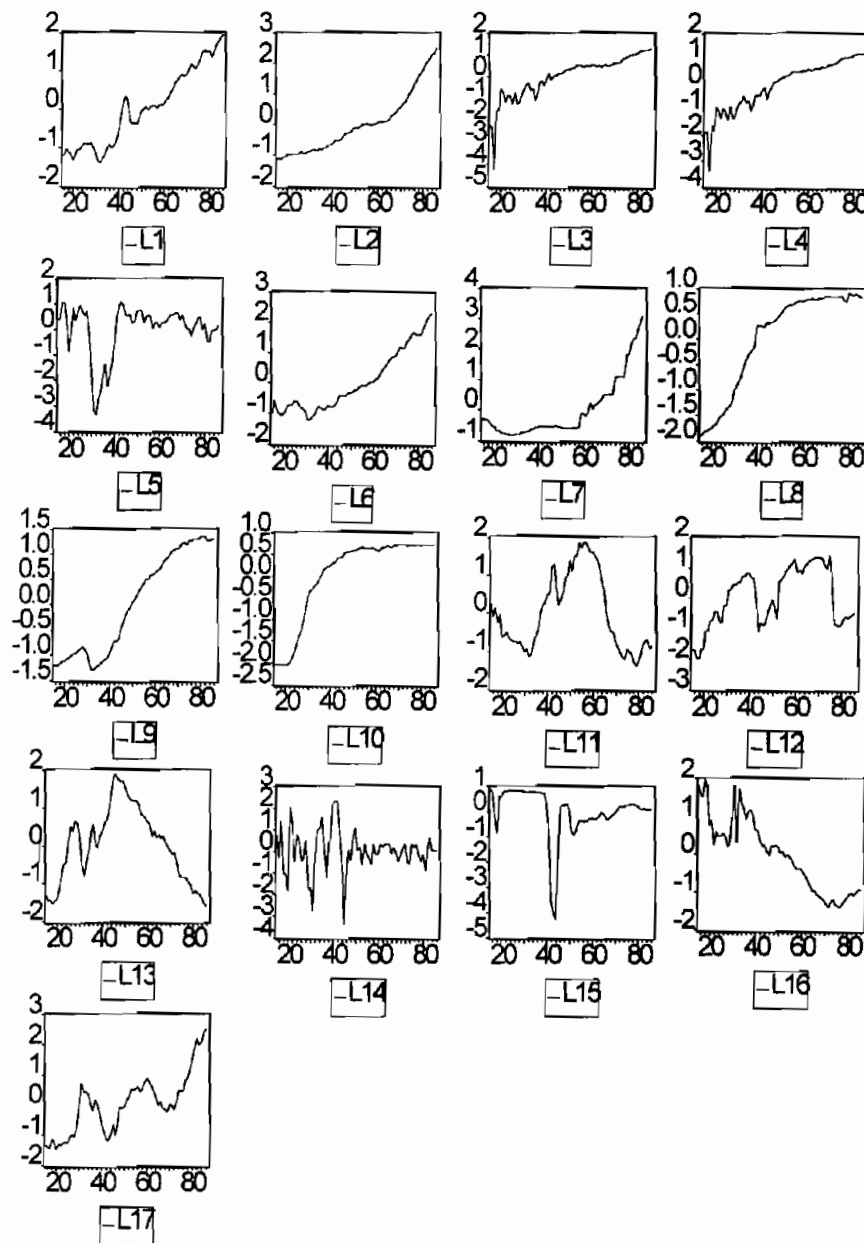


Fig. 1. Time plots of the series, (mean=0; variance=1).

models. This is only necessary to provide estimates of the innovations of the series, which are then analyzed, for their clustering properties. In addition, the method we use does not violate the triangle inequality while at the same time it incorporates the volatility of the series without the need for the reliance on the assumption of asymptotic normality needed in Maharaj's method.

This analysis can be viewed as similar to principal component analysis in which the correlation matrix of these variables is used to define the interrelationships between the series. However, it is widely recognized that correlations are prone to spurious inferences especially when the series are subject to unit roots. From the plots in Fig. 1 one can see that many of these annual series

seem to possess the characteristics that would indicate they could be modelled as first-order autoregressive processes with a parameter of one. In formal tests for unit roots (using the Augmented Dickey–Fuller test with two lagged values of the difference and an intercept), we find that only in L3, L4 (the two life expectancies), L14 (the rate of GNP growth) and, L15 (the percentage of GNP not for defence) could we reject the hypothesis of a unit root at the 5% level of significance. Thus, the use of the standard methods that rely on functions of the levels directly will lead to inferences that are based on the unit roots that underlie these series. The alternative, to use only the changes in the series, results in the removal of the important information that is contained in the levels.

The information based distance metric proposed in Hirschberg et al. (1998) is a method for combining both the locations of the series and the variation in these series without the violation of the triangle inequality and furnishes a method for the comparison of stationary and nonstationary series.

The clustering method used in this analysis is a hierarchical agglomerative clustering which uses the average linkage to determine intercluster distances. This method clusters up from the case where every series is in a cluster of its own to the case where they are in one cluster. The movement into a cluster is determined by the average distance that a candidate member is to the all the existing members of the cluster considered for membership. This process also applies to combining clusters by determining the distance as the average distance between all members of the two clusters considered for membership in the new single cluster.

The distinguishing feature of the present application, over the use of traditional cluster methods that are available in standard statistical packages as documented in Afifi and Clark (1996), is the use of an information metric to define the distance. The distance metric employed here is based on the Bhattacharyya (1943) and Matusita (1967) affinity measure between two distributions (ρ_{ij}^*) (see Maasoumi, 1993 for details) which is defined as:

$$\rho_{ij}^* = \int_{-\infty}^{\infty} f_i(x)^{1/2} f_j(x)^{1/2} dx$$

where $f_i(x)$ and $f_j(x)$ are the densities of the two measures being compared. The information distance measure is given by $1 - \rho_{ij}^*$.

In the present case the simple comparisons of the density based on the annual observations for each measure from 1915 to 1987 would not account for the time related dimension of each attribute's time series. It is possible that two attributes' time series will have very similar densities but that they are independent across time. Thus to form the density estimates we employ a two step process. In the first, we establish a density for the innovation in each attribute's time series. In the second step, we locate the density so that comparisons can be made for each year. This is done by fitting a set of time series models to each attribute's time series after each indicator has been scaled with a mean of zero and a variance of 1.

Each series was modelled as either an ARIMA(1,1,1) or an ARMA(1,1) depending on whether we have determined the presence of a unit root. These models are constructed purely to provide estimate of the innovations, or time series shocks, to which these series are subjected. These models were found sufficient to remove the autocorrelation from the estimated errors. Thus once we have determined the degree on integration of the ser-

ies we then perform a parametric estimation in which the series are fit to one of two models. In order to account for the possible implications of a shock due to the Second World War we have also estimated an intercept shifter for those years. Once this modelling exercise is done, we have two series for each variable. Fig. 1 shows the time plots of the levels of each series. We also were unable to reject the null hypothesis of any remaining autoregressive structure in the residuals. The next step is to use the residuals as estimates of the innovations of the time series to estimate a nonparametric density.

The density estimates were constructed using a normal kernel and 1/4 the window width specification recommended by Scott (1979).⁴ The estimated densities are shown in Fig. 2. Note that these plots have vertical axes that are automatically scaled, however, some of these distributions are much more diffuse than others. For example, the GNP growth rate (L14) has a very high variance and a marked degree of asymmetry compared to the percentage of homes with phones (L9).

A distance ($D(i,j)$) is then computed for each pair of series by locating the mean of the estimated distribution of the innovations to the value of the series for that year and summing them over all the years in the sample. Thus even though two time series may move together if one is subject to a greater level of variability then they will not necessarily be considered to be closer to each other than to some other series for which the innovations are less variable. Formally the distance measure ($D(i,j)$) is the average of the information for the entire series.

$$D(i,j) = 1 - \frac{1}{T} \sum_{t=1}^T \left[\int_{-\infty}^{\infty} \hat{f}_i(x+z_{it})^{1/2} \hat{f}_j(x+z_{jt})^{1/2} dx \right]$$

where the estimated densities for the innovations are shifted to account for each observation of the series (z_{it} and z_{jt}) and $T=73$.

Table 2 provides the values of the distance between each attribute's time series (scaled from 0 to 100). Once the distances are computed, we can then find the two series that are closest (in this case female and male life expectancy with a distance of 15). The distance to the clusters is then determined by the use of the average linkage method which computes the average value of $D(i,j)$ between each possible member of the proposed cluster to determine cluster membership. Note that the values on the last two rows indicate the distance of SO₂ and NO_x from the other series. Thus, we see that many of the other series are closer to the pollution control mea-

⁴ The window or bandwidth $h=(3.49sdT)^{1/3}$ where sd is the standard deviation and T is the number of observations. This bandwidth often resulted in an overly smooth density for these series so we used $h/4$ as the bandwidth here.

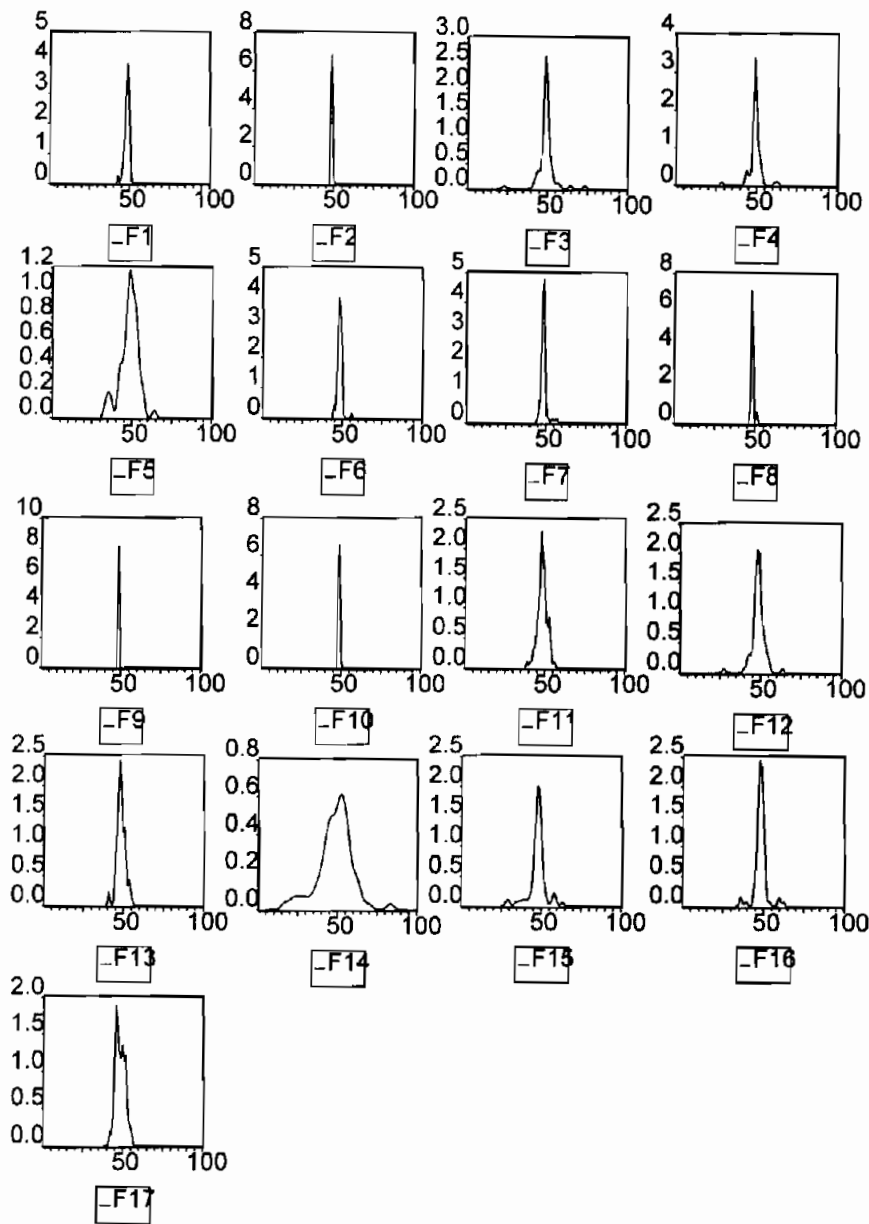


Fig. 2. Estimated densities of the estimated innovations using a normal kernel. The expected value of the innovations is 50 for these plots.

tures than they are to each other. We see that the control of NO_x is closer to GNP rate of growth than to any other series. In addition, that control of SO_2 is closest to male life expectancy (L3) than to any other series. Interestingly, the series for the control of SO_2 is more closely related to all the other series in this analysis than to the control of NO_x .

The next aspect of cluster analysis is the determination of how many clusters best represent the interrelationships in the data. In the case of principal component analysis one would plot the value of the eigen values by the number of components where the components are ordered by the size of the principal components in order to form a downward sloping “scree plot”. In the hierarchical cluster analysis, the equivalent plot is one of dis-

tance necessary to combine the last cluster on the vertical axis with the number of clusters on the horizontal axis. Thus the distance that must be tolerated to reduce 17 clusters to 16 clusters is the smallest, and the distance to combine two clusters to one cluster (the cluster formed by all the series) is the greatest. This plot can then be used in the same manner as the scree plot for principal components as a means of determining the minimum number of clusters that can be formed before the differences between the series are too large. Mojena (1977) suggests that the determination of how large is too large be based on a formal test of the deviation of a distance from the mean of the distances when they are normally distributed. It appears that there is no theoretical justification for this test. Thus in this analysis we will show

Table 2
 $D(i,j)$ for the series

Series	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12	L13	L14	L15	L16	L17
Per capita GNP	L1	0	61	58	73	39	72	84	62	89	85	70	85	68	84	85	64
1/infant mortality	L2	61	0	75	76	64	63	95	80	96	89	86	89	80	87	91	75
Male life expectancy	L3	58	75	0	15	51	66	54	72	60	63	54	64	48	67	77	49
Female life expectancy	L4	58	76	15	0	58	66	65	67	65	70	57	72	56	75	83	54
Employment rate	L5	73	85	51	58	0	75	73	76	84	77	62	62	63	31	44	63
Per capita Disp inc	L6	39	64	66	66	75	0	60	90	71	92	80	71	86	69	80	88
Physicians/100,000	L7	72	63	62	65	73	60	0	89	87	94	74	76	83	66	71	87
llyw miles per capita	L8	84	95	54	50	76	90	89	0	91	63	82	75	84	75	86	91
Homes with phones (%)	L9	62	80	72	67	84	71	87	91	0	91	90	78	91	82	92	94
Homes with radio (%)	L10	89	96	60	65	77	92	94	63	91	0	88	81	85	78	89	94
100,000/homicide	L11	85	89	63	70	62	80	74	82	90	88	0	63	58	55	73	69
Percentage age 5–17 in school	L12	70	86	54	57	62	71	76	75	78	81	63	0	61	50	65	70
Newspapers/cap	L13	85	89	64	72	63	86	83	84	91	85	58	61	0	57	70	72
GNP rate of growth	L14	68	80	48	56	31	69	66	75	82	78	55	50	57	0	35	52
Percentage of GNP not defense	L15	84	87	67	75	44	80	71	86	92	89	73	65	70	35	0	52
Population/NO _x	L16	85	91	77	83	63	88	87	91	94	94	69	70	72	52	52	0
Population/SO ₂	L17	64	75	49	54	64	67	68	74	75	79	77	54	64	54	63	84

all the cluster memberships as they are formed and examine a particular set of clusters based on the indications of the plot of the distances to combine versus the number of clusters.

4. Clustering results

The results of the cluster analysis are summarized in the dendrograms in Figs. 3 and 4. In Fig. 3 the number of clusters is on the horizontal axis and the cluster, membership is on the vertical axis. This diagram can also be used to investigate the genealogy of any of the clusters. In addition, this diagram demonstrates a potential problem with agglomeration clustering — that once the clustering algorithm puts an individual attribute in a cluster it will never be removed and put into another. Fig. 4 is another version of the dendrogram however in this case the distance is the unit of measure on the horizontal axis. From this plot, we can see that the distances between most series that are included in the same cluster lie in

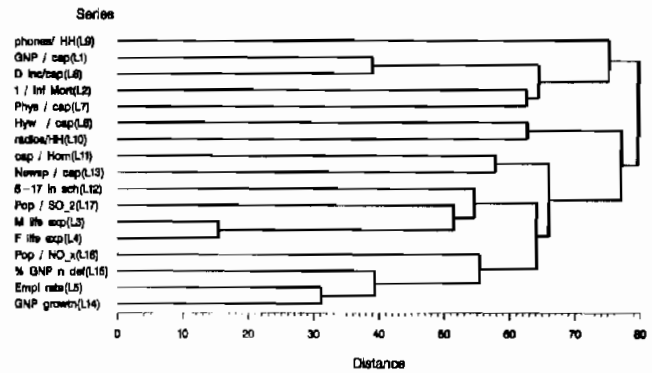


Fig. 4. Dendrogram by distances based on $D(i,j)$.

a range of 14–66. Note that Fig. 4 shows that the intracluster distances up to the case of four clusters are relatively low. However, as we move to combine the series into three or fewer clusters we are combining series that are further and further apart.

Fig. 5 is the plot of distances that can be used to deter-

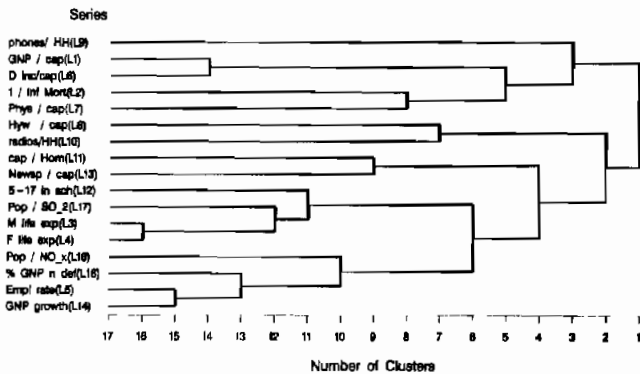


Fig. 3. Dendrogram by number of clusters based on $D(i,j)$.

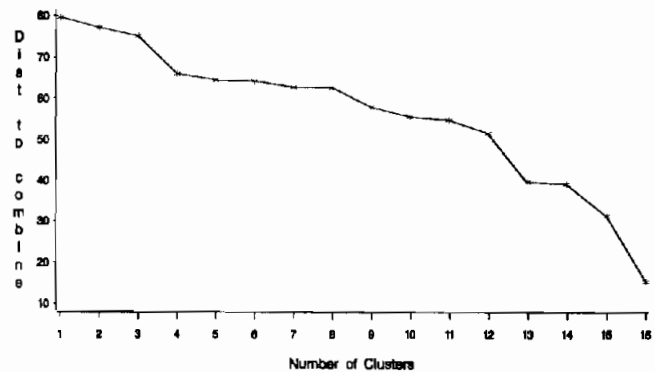


Fig. 5. Distances to combine each cluster.



Fig. 6. Change in distance to combine from last cluster.

mine the appropriate number of clusters. From this plot, it can be noted that the change in the distance from 3 to 4 shows a definite shift in the angle. Fig. 6 plots the changes in the distances. More dramatic distances/dissimilarities must be tolerated in pushing attributes from 4 to 3 clusters. This would seem to be a point at which distinct dimensions of quality are identified. Table 3 below indicates the membership in four clusters.

The membership of cluster #2 indicates that air quality measures are statistically similar to most of the life expectancy indicators, as well as some of the other non-macro welfare attributes. The exceptions are infant mortality which appears to be generally closely represented by rich/poor economic attributes (see also Hirschberg et al. (1991, 1998), and the distinct members of the clusters # 3 and 4 which indicate communications attributes deserve distinct representation in multidimensional considerations.

Table 3
Membership in the four clusters

<i>Cluster #1</i>	
GNP per capita (L1)	
1/Infant mortality (L2)	
Disposable income per capita (L6)	
Physicians per capita (L7)	
<i>Cluster #2</i>	
Male life expectancy (L3)	
Female life expectancy (L4)	
Employment rate (L5)	
Population per homicide (L11)	
Percentage of children 5–17 in school (L12)	
Newspapers per capita (L13)	
GNP growth (L14)	
Percentage of GNP not for defence (L15)	
Population/NO _x (L16)	
Population/SO ₂ (L17)	
<i>Cluster #3</i>	
Highway miles per capita (L8)	
Radios per household (L10)	
<i>Cluster #4</i>	
Phones per household (L9)	

5. Conclusions

It is worth noting that the similarity between the air quality indicators and many of the other members of cluster #2 becomes evident quite early in the clustering search. In fact as soon as 10 clusters are formed, the above mentioned similarity is revealed. Hence we may safely conclude that our main result, being based on very effective measures of information, is very strong. The control of SO₂ remains very closely related to the life expectancy series and the control of NO_x is more closely related to the three macroeconomic variables. This would possibly indicate that economic factors allowed the change in NO_x production but health concerns were more involved in the control of SO₂.

To explore other aspects of our series we fit two simple regressions (which are subject to the problem of spurious regression), in which these pollution controls are the dependent variable and all the other variables are used as regressors. In the model for SO₂ we could not reject the hypothesis that the parameters for the two life expectancy series were zero. In addition, in the model for NO_x we found no evidence for a relationship with any of the macro economic variables. Given the non-stationary nature of most of these series this is not a surprising result but it does highlight the shortcomings of solely relying on correlation analysis.

In sum, we find that clusters that include the pollution indices may be employed by investigators to represent dimensions of welfare that include, as well as go beyond, these simple indices.

Acknowledgements

We would like to thank Michael McAleer for his comments on an earlier version of this paper. Any remaining errors are our own.

References

- Bhattacharyya, A., 1943. On a measure of divergence between two statistical populations defined by their population distributions. *Bulletin Calcutta Mathematical Society* 35, 99–109.
- Bresnahan, B.W., Mark, D., Shelby, G., 1997. Averting behavior and urban air pollution. *Land Economics* 73, 340–357.
- Kaufman, L., Rousseeuw, P.J., 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York.
- Hirschberg, J.G., Esfandiari, M., Slottje, D.J., 1991. Cluster analysis for measuring welfare and quality of life across countries. *Journal of Econometrics* 50, 131–150.
- Hirschberg, J.G., E. Maasoumi, Slottje, D.J., 1998. A cluster analysis the quality of life in the United States over time, Department of Economics research paper #596, University of Melbourne, Parkville, Australia, www.ecom.unimelb.edu.au/ecowww/research/596.pdf.
- Maasoumi, E., 1986. The measurement and decomposition of multidimensional inequality. *Econometrica* 54, 991–997.

- Maasoumi, E., 1993. A compendium of information theory in economics and econometrics. *Econometric Reviews* 12, 137–181.
- Maasoumi, E., 1998. Multidimensional approaches to welfare. In: Silber, L. (Ed.), *Income Inequality Measurement: From Theory to Practice*. Kluwer, New York.
- Maharaj, E.A., 1995. A significance test for classifying ARMA models. In: *Proceedings of the 1995 Econometrics Conference at Monash*, Monash University, Department of Economics, pp. 219–251.
- Matusita, K., 1967. On the notion of decision functions. *Annals of the Institute of Statistical Mathematics* 19, 181–192.
- Mojena, R., 1977. Hierarchical grouping methods and stopping rules: an evaluation. *Computer Journal* 20, 359–363.
- Piccolo, D., 1990. A distance measure for classifying ARIMA models. *Journal of Time Series* 11, 153–164.
- Scott, D.W., 1979. On optimal and data-based histograms. *Biometrika* 66, 605–610.
- Slotje, D.J., Scully, G.W., Hirschberg, J.G., Hayes, K.J., 1991. *Measuring the Quality of Life Across Countries: A Multidimensional Analysis*. Westview Press, Boulder, CO.
- US Environmental Protection Agency, 1998a. *National Air Pollutant Emission Trends Report, 1900–1996*. Office of Air Quality, Planning and Standards, Research Triangle park, NC, 27711, www.epa.gov/oar/emtrnd.
- US Environmental Protection Agency, 1998b. *National Air Pollutant Emission Trends, Procedures Document, 1900–1996*. Office of Air Quality, Planning and Standards, Research Triangle park, NC, 27711. www.epa.gov/oar/emtrnd.