

Clusters of Attributes and Well-Being in the US

Joseph G. Hirschberg¹, Esfandiar Maasoumi and Daniel J. Slottje²

February 20, 2001

ABSTRACT

Using ARIMA models and entropy, the dynamic evolution of several functions of aggregate income and other attributes of well-being is analyzed for statistical "similarity" in order to determine potentially distinct dimensions in multidimensional analysis of welfare and quality of life in the US. The entropy metric compares entire distributions and is more general than principal components and other correlation-based techniques for clustering. To help macroeconomic policy makers, we compare the distribution of several composite measures of well-being which include income, with the distribution of some common measures of aggregate income over the period 1915-1995. Per capita disposable income and growth in GNP are statistically distinct dimensions of wellbeing.

Key Words: Time series, Information measures, Well-being, Entropy, ARIMA, Cluster analysis.

JEL classification: C82, I31, C14

1 Introduction

Much of modern empirical macroeconomics is concerned with time series models of the conditional mean or variance of income. Less attention is paid to the whole distribution of income (or other variables) from which our observations may have been drawn. The welfare basis of economics requires that we learn about the determination of income (and employment), and its distribution. Some early work developed diffusion processes which produced steady state distributions for income or wealth. Examples of this include Champernowne (1953), Gibrat (1931), Rutherford (1955) and Sargan (1957). Sargan considered a relationship representing the effects of four separate causes of change in wealth which will explain its evolution over time. This was a causal extension of the model suggested by Champernowne for income. It produced differential equations whose steady state solutions have provided durable support for log normal and Pareto distributions in macroeconomics. Log normal is also a workhorse in finance. Similarly, Langley (1950) had analyzed the distribution of private capital.

¹Dept. of Economics, University of Melbourne, Parkville, 3052, Australia, (03) 9344-5273.

²Dept. of Economics, Southern Methodist University, Dallas, TX 75275. Corresponding Author: maasoumi@mail.smu.edu. This paper is dedicated to the memory of Denis Sargan as it deals with fundamental issues that arise in a multidimensional extension of his work on the evolution of wealth distribution in Sargan (1957). We thank without implicating, Tom Fomby, David Hendry, Hashem Pesaran, several referees, and participants at our seminars.

There are other important attributes that impinge upon well-being, such as health, education, environment, and freedoms. In recent times attempts have been made to formally acknowledge that well-being is a function of several arguments, including income and the GNP. Sen (1985, 1987) has examined this question. He points out that capabilities are as important as commodities in analyzing the quality of life. Thus, examining only income or only expenditures as arguments in a utility or social welfare function may be inadequate.³ Sargan and his contemporaries did not have access to data on other indicators of well-being. We do, and well-being is the central question of macroeconomic policy that requires a better understanding of the dynamic evolution of income as well as other indicators.

In this paper we compare the whole distribution of income with the marginal distributions of other indicators of wellbeing, as well as several composite indices of well-being. In statistics, two variables are in general identical/similar if their distributions (or characteristic functions) are identical/similar, not just their means or a few attributes of their distributions. Comparison of distributions is generally incomplete when only moments and their functions are compared. We use entropy measures of distance between whole distributions, and we avoid a priori assumptions about the family of distributions and use nonparametric density estimators instead.

There is now a rather extensive literature on the relationship between "growth" (of incomes) and "life". One interpretation of this literature is that it attempts to reduce all attribute dimensions to that of income (GNP). Easterly (1997) provides a critical survey of the largely cross section evidence in favor of such reductions. His own panel data model for 95 "life" indicators, controlling for country specific factors, finds scant supportive evidence for a positive and significant relation between growth and most measures of well-being. This indicates the existence of distinct dimensions to welfare. But this type of similarity analysis is limited to correlation measures. For non-normal and/or nonlinear processes, it can be very incomplete or even misleading. Also, 95 indicators will surely lead to some double counting of similar welfare dimensions.

When data are available on desired attributes, the identification of distinct attributes is a statistical question. The question is whether a candidate attribute adds to the appropriate statistical information set. This question is generally too broad to be determined by merely the correlation between variables, or only by the integration/cointegration properties of attributes. In statistics, two variables are in general identical/similar if their distributions (or characteristic functions) are identical/similar, not just their means or a few attributes of their distributions. Comparison of distributions is generally incomplete when only moments and their functions are compared. We use entropy measures

³ Kolm (1977), Sen (1985, 1987), Atkinson and Bourguignon (1982), Maasoumi (1986) and others have discussed these issues. In recent years there have been several attempts at practical implementation of this philosophy, aimed generally at providing summary indices of well-being. Also, major attempts at compilation of data, such as the United Nation supported Basic Needs (BN) and Physical Quality of Life Indicators (PQLI), may be cited; see Ram (1984).

of distance between whole distributions, and we avoid a priori assumptions about the family of distributions and use nonparametric density estimators instead. But moments-based and whole distribution comparison techniques can be complementary and confirmatory. For instance, processes may be conditionally different in their means and very similar or identical in their innovations. Thus, we first fit ARIMA models to the means, and proceed to evaluate similarity or entropy distance on the basis of the whole distributions. Reassuringly, our entropy measure reduces to increasing functions of the correlation measure for jointly normally distributed variables.

We analyze fifteen apparently distinct variables indicating well-being and quality of life in the US. We use a transformation of the Matusida-Bhattacharya-Hellinger entropy metric to measure the "distance" between attributes in order to identify distinct clusters of attributes. Other entropy measures (including Shannon's) record "divergence" and violate the triangle rule. Inconsistent decisions can then result in determining cluster membership for more than two variables. In our approach, we first fit univariate time series models to each series. Any similarity in the conditional means is revealed in this first step. We then go beyond and estimate non-parametric kernel densities of the residuals of each attribute obtained from the first step. By centering each series at their observed values the conditional mean information is preserved. Finally we measure the entropy distances between the entire distributions in order to determine distinct clusters. In this way we are able to present a decomposition of the distance between these attributes into two components. One is the entropy distance between the non parametric residual distributions. The second is due to the mean differences over time.

Most of the work that has been done on analyzing the quality of life within a country has focused on comparisons of urban locations using hedonic price models, cf. Rosen (1974,1979), Roback (1980,1982), Blomquist et al. (1988) and Gyourko and Tracy (1991) for examples of work in this vein. Nordhaus and Tobin (1972) created a measure of economic welfare which adjusted GNP for leisure, women engaged in household production and urban disamenities. Their measure still focused on GNP, however. Slesnick (1991) discussed the idea of a standard-of-living index based on aggregate expenditure per capita. Other clustering analyses have usually been based on cross section data. Here we analyze income and clusters of quality of life indicators in a particular geographic region over a relatively long time period ⁴.

By uncovering similar attributes and thereby allowing a reduction in dimension, a related problem is also lessened. This is the difficult and subjective question of how to weight distinct attributes (i.e., choosing a cardinal welfare

⁴ Slesnick (1991) analyzed living standards over time. He relied on expenditure surveys but did not have contiguous data nor did he go back in time as we do here. Ram (1984) suggested Principal Components (PC) of the PQLI and the BN data, Maasoumi and Jeong (1985) and Maasoumi (1989) proposed information theoretic indices of the same data as Ram, Slotje (1991) studied several indexing techniques based on hedonic regressions, PCs, and "ranked attributes", using many more economic and social attributes than usual, and Hirschberg et al (1991) studied economic and social indicators, including civil liberties and labor force participation in a cross section of international data.

function). It is relatively easier to attach weights to a few clusters of attributes than many of their constituent variables. It is also easier to conduct sensitivity studies that determine the robustness of any qualitative inferences within sensible bounds for these weights.

Our findings reveal that two important thresholds must be crossed in reducing welfare dimensionality. One is going from 15 dimensions to 10, the other is in reducing to fewer than four clusters. Relatively little distance needs to be tolerated to reduce ten clusters to four. We think there are at least four distinct dimensions in these 15 attributes. "Income" may represent only one of these four dimensions/clusters. Interestingly, the Growth rate of GNP (while volatile) is on average the closest to all other attributes, and represents its own cluster well. But it is a statistically distinct dimension to that of per capita Disposable Income (well representing its own cluster).

The plan of this paper is as follows. Section two describes the entropy based method of cluster analysis employed here. In section three the 15 attributes and our data sources are described, and section four reports our empirical findings. We provide graphs in which the evolution of national income is compared with that of several aggregate measures of well-being. Section five concludes the study.

2 Cluster Analysis

The concept of clustering time series has been dealt with by a number of authors. An early paper in this area is by McGee and Carlton (1970) in which they investigate the problem of defining clusters of observations in univariate time series in order to detect regime changes. More recently Piccolo (1990) and Maharaj (1995) have proposed grouping time series based on the parameters of an ARIMA model fitted to the time series. Piccolo looks only at the parameter estimates while Maharaj uses the Chi square distributed distance as employed in Hirschberg and Dayton (1996) for clustering regression parameters based on the asymptotic distribution of the estimated ARIMA parameters. In a recent issue of this Journal, Hobijn and Franses (2000) analyze the time series properties of per capita productivity for 112 countries over 29 years. Their goal is to establish which economies are converging to "clubs" of similar productivity growth. One difficulty with these studies is the implicit assumption that a parameterized (conditional expectation) model explains fully the distribution of the variables under review. In the present case 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 distance used by Maharaj and others violates the triangle inequality. 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.

In our approach the focal point is the entire distribution of the attribute over time which we estimate in a nonparametric manner. Clusters are formed by groups of attributes that are most "similar", according to our metric entropy.

An analysis could be based on the correlation matrix and its eigen vectors using techniques related to principle component analysis and factor analysis. These methods generally allocate an equal weight to each attribute and are notoriously prone to outlier influences. We believe it can be misleading to limit our view of variables and their relations to correlations.

The Bhattacharyya (1943) and Matusita (1967) entropy affinity measure between two distributions ($\frac{1}{2}i_j$) is defined as:

$$0 \cdot \frac{1}{2}i_j = \int_{-\infty}^{\infty} f_i(x)^{1/2} f_j(x)^{1/2} dx$$

where $f_i(x)$ and $f_j(x)$ are the densities of the two attributes being compared. Then our basic distance measure is :

$$D(i;j) = 1 - \frac{1}{2}i_j = \frac{1}{2} \int_{-\infty}^{\infty} (f_i^{1/2} - f_j^{1/2})^2 dx$$

It is zero if the two densities are identical. This measure is a component of the Hellinger distance that can be used to measure and test for goodness of fit, serial dependence of possibly nonlinear variables (Skaug and Tjostheim (1996), Granger et al. (1999)), and predictability (Maasoumi and Racine (2001)). It is of some importance to appreciate that, while the choice of distance measures is similarly as difficult as cardinalization of utility functions, it is not as arbitrary as common practice would suggest. There are axiom/property systems that can derive divergence/entropy/utility measures which would eliminate many candidates and obtain "ideal" measures. Entropies, especially the Bhattacharyya-Matusida one, are obtained in this way. This is discussed in detail in Maasoumi (1993). Granger et al. (1999) and Skaug and Tjostheim (1996) enumerate several important properties for the entropy metric used here. Importantly, these include applicability to both continuous and discrete variables, invariance to nonlinear transformations, metricness, and reduction to increasing functions of the correlation coefficient. The asymptotic and bootstrap distributions are discussed and implemented for time series in the last two citations. Correlation and its functions dominate the cluster analysis field. Instead, our metric is a generalization which is also robust to distribution assumptions, possible nonlinearities, and discreteness.

We estimate a density for the innovations in each attribute's time series. This is done after fitting a set of time series models to each attribute (each has been scaled to a mean of 0 and a variance of 1). The residual densities can be very similar even for independent innovations. But the attributes can have trending behavior and conditional means that are similar or completely unrelated. But we re-center the distributions on the observed data values at each point in time and compute the corresponding $D(i;j)$ for the entire sample. This preserves any long run information in the series. The fitted models were either ARMA(1,1) or ARIMA(1,1,1) depending on the potential presence of a unit root in the series.

$$x_{i;t} = a_i x_{i;t-1} + d_i w_{i;t} + b_i^2 x_{i;t-1}^2$$

or, $\Phi x_{i;t} = a_i \Phi x_{i;t-1} + d_i w_{i;t} + b_i^2 x_{i;t-1}^2$

where a , b , and d are the parameters which we estimate, $x_{i;t}$ is the series, $\epsilon_{i;t}$ is the random error, and war_t is a dummy variable for the years during the second world war. These models are found to be the most parsimonious models that capture the conditional means of these series.

Except for L5 (the employment rate), L13 (the newspaper circulation per capita), L14 (the rate of GNP growth) and L15 (the % of GNP not for defence), we could not reject the hypothesis of a unit root for all the other series. See the next section for fuller definitions of all the variables. In each case we used the Portmanteau (Q) test for randomness of the estimated residuals from the model and found that we could not reject the null hypothesis of randomness at better than a 90% level in almost all cases. But as demonstrated in recent work, such tests cannot detect nonlinear dependence. See Granger et al. (1999) and their references. We do not conclude that these residuals are "similar".

Using a normal kernel and 1/4 the window width specification recommended by Scott (1979),⁵ we estimated a density function for residuals of each series⁶ with each density evaluated at 1000 equally spaced points⁷. Then, we compute the average of the $D(i; j)$ s over the entire sample. This metric not only compares the values of the attribute's time series as it shifts over time but it also incorporates additional distributional information in the innovations. Even though two attribute's time series may move together, if they have different innovation densities (shocks or volatilities, for instance), they may not be considered to be "close" to each other. The overall distance measure $D(i; j)$ is estimated as the scaled average of our entropy distance for the entire sample:

$$\bar{D}(i; j) = \frac{1}{T} \sum_{t=1}^T \int_{-\infty}^{\infty} f_{it}(x)^{1/2} f_{jt}(x)^{1/2} dx$$

where $f_{it}(x)$ is the estimated density for the innovations for series i located at the observation for time t , $f_{jt}(x)$ is the corresponding estimated density for series j at time t , and $T = 81$.

The method we use falls under the general class of hierarchical agglomerative clustering techniques. To begin with, each attribute is the only member of one cluster. Then, using some measure of fusion or distance, each cluster is considered for association (clustering) with every other cluster, in successive stages. At first, 15 clusters/dimensions are available, then 14, 13, . . . , etc., until there remains one cluster with all the attributes in it.

⁵ The window or band width (h) is $h = 3.49 \text{ sd } n^{-1/3}$; where sd is the standard deviation and n 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. Other bandwidths were used such as the one found by Bowman (1985) $h = 2 \text{ sd } n^{-1/5}$ however they resulted in no difference in the clustering results reported here.

⁶ Other kernels were tried including the Epanechnikov (1969), the triangular, and the rectangular with very little variation in the resulting distances computed.

⁷ The number of points used did not have an effect on the clustering when at least 200 were used. The horizontal scale in these figures is the same and the density plots are centered at 500. The vertical axis in every case is scaled differently so that the integral of each density is equal one.

Table 1 reports the overall distances between the attributes. The permissible range is 0 to 100. We see that the two closest series are female and male life expectancy (L4 and L5) with a distance of 15. These two are combined to form one cluster (the 14th). The other thirteen attributes/clusters now considered as candidates for further clustering. To form 13 clusters we either add another variable to the two life expectancies to form a new cluster with three members, or we form a new cluster consisting of two or more of the remaining series. In this case we found that the employment rate (L5) and the GNP growth rate (L14) are the closest to each other with a distance of 32 and they are combined to form a cluster. We use the average linkage method which computes the average value of $\bar{D}(i;j)$ between each candidate attribute and the members of the existing clusters. The attribute(s) with the smallest such distance joins in an existing cluster. If none is closer to an existing cluster than to some of the remaining nonclustered variables, closest combinations amongst the latter will form new clusters. It is here that violation of the triangle rule by a nonmetric measure can lead to inconsistent decisions.

	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12	L13	L14	L15	
L1		66	58	57	78	39	77	88	69	90	91	80	88	71	83	L1
L2			75	77	89	63	58	95	91	97	91	90	91	82	84	L2
L3				15	55	66	68	65	66	71	74	65	74	50	67	L3
L4					59	66	73	61	61	69	78	65	79	55	74	L4
L5						82	81	76	83	76	64	61	65	32	43	L5
L6							62	92	76	93	87	82	88	73	79	L6
L7								93	91	94	81	84	89	71	71	L7
L8									91	60	88	81	91	76	86	L8
L9										93	94	82	93	80	90	L9
L10											89	84	89	78	88	L10
L11												63	63	56	73	L11
L12													68	49	63	L12
L13														60	72	L13
L14															35	L14
L15																L15

Table 1: Distances $D(i;j)$ computed between each series.

The distance that must be tolerated for a new entry into the existing clusters continues to grow larger. An important aspect of the hierarchical clustering is the determination of the appropriate number of clusters. This question has been the subject of a number of studies that deal with this "stopping rule problem". Here we have adopted the same kind of methods used in factor analysis based on the "scree" plots. These normally plot the eigen values of variables for "elbow points". Mojena (1977) suggests a similar method for the examination of the distances needed to form successive clusters. Accordingly, we plot the change in the distances between the variables that formed a given cluster at each stage⁸.

⁸The decision to stop may be made on the basis of the statistical significance of the dis-

3 Attributes of Economic Well-Being

Since our objective is to measure well being as comprehensively as possible, a total of 17 indicators of the quality of life were selected. The attributes are from a number of sources which are listed in the appendix. The time series are for the years 1915-1995. One important selection criterion was that the observed series went back sufficiently long enough in time. The plot of each time series is given in Figure 1. All of these attributes have been rescaled to have a mean of zero and a variance of one, and the last two digits of the year are on the horizontal axis.

Attribute L1 is annual per capita Gross Domestic Product of the United States in real 1958 dollars. Per capita real GDP is one of the best representations of the command over resources and the macroeconomy. All of the attributes that are measured in dollar terms have been deflated by the GNP deflator to put them in real terms and in per capita terms where appropriate. This is necessary to eliminate the momentum effects which would always give later years higher values and consequently higher rankings. We adopt the point of view that per capita GDP contributes positively to well being.

L2 is the inverse of the Infant Mortality Rate (IMR). This variable is included because it may be interpreted as a proxy for deprivation suffered by those at the lower end of the wealth distribution. It also serves as a painful reminder that while the U.S. may have a powerful economy in other ways, it also has one of the highest IMR's among developed western countries.

Attributes L3 and L4 are the life expectancies for males and females, respectively. Life expectancy is associated with better medical, sanitary, and other quality of life conditions. We are aware that this view is not universally held.

The employment rate is L5 (1 minus the unemployment rate) and mean income per household is L6. Persistent unemployment is highly correlated with poverty and many other social ills. Mean real income per household is an indicator of the household unit's control over resources. It is assumed that as both L5 and L6 increase, the quality of life increases. L7 is the number of physicians for every 1,000,000 people in the population. The eighth attribute, L8, is the total number of rural and urban federal highway miles per capita. L9 and L10 are the number of telephones per 1000 people, and the total number of households with radio receivers which are more of interest in a historical context than for predictive purposes since both technologies have long since reached almost 100% saturation. These attributes are included to capture the quality of health care, the ability to be mobile, and the ability to communicate at the simplest level. It is assumed that all four indicators are positively related to the living standard. The inverse murder rate and % of children ages 5 to 17 enrolled in school are attributes L11 and L12. An environment that is stressful

tances. In related work with a normalized version of our entropy measure, Granger, Maasoumi, and Racine (1999) find that it has predictably large bootstrapped standard errors. Density estimators tend to exhibit this behaviour. Such a bootstrap study is beyond the scope of this paper. But our experience suggests that the major kinks in the scree diagrams correspond to the likely significant changes in the distances.

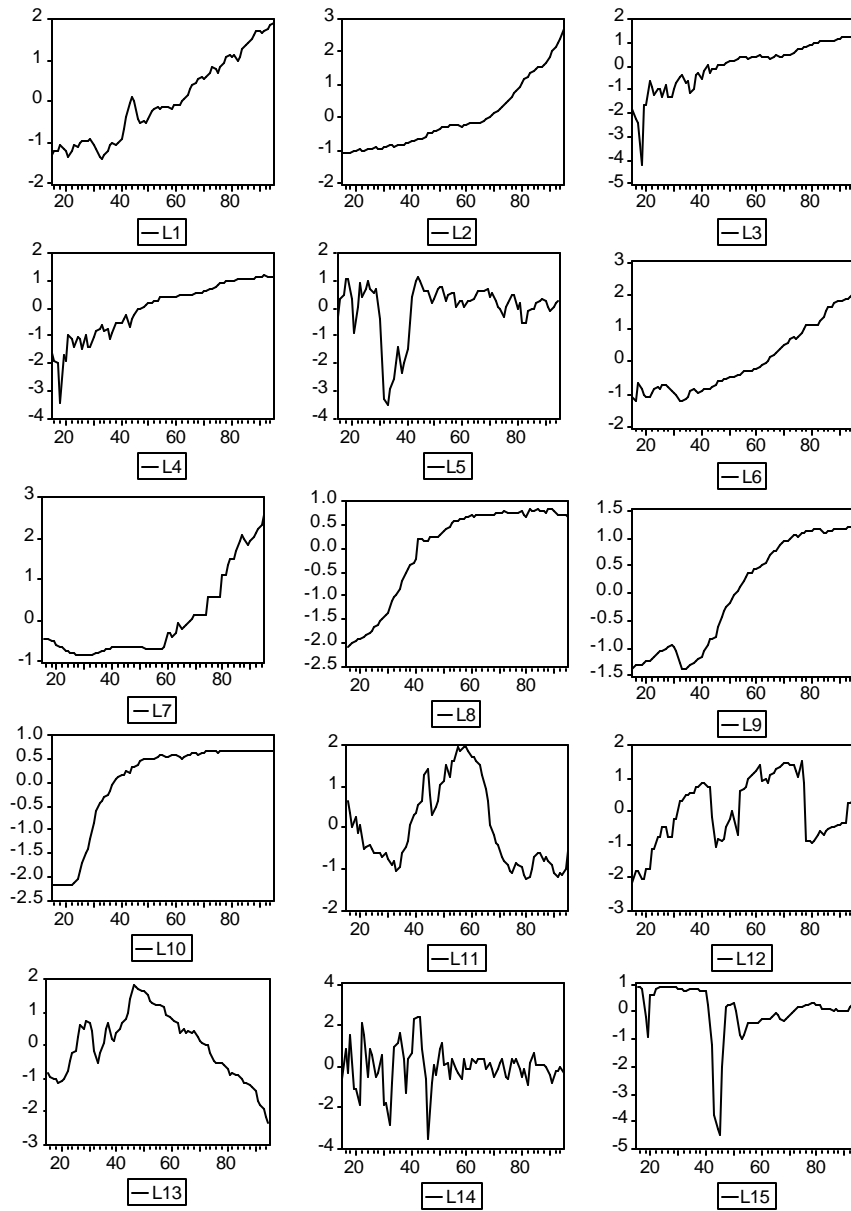


Figure 1: The time series (mean=0, and variance=1)

from fear of attack is less desirable. Dropout rates and average number of years of schooling completed may be more appropriate than L12, but the data are not available back to 1915. Higher levels of schooling suggest an increased ability to enjoy life. Increased schooling also has positive externalities such as lowering the crime rate and the unemployment rate.

The circulation of daily newspapers per capita in the United States is L13. It is included as an indicator of access to public information and as an indicator of the ease of acquiring this information. This variable is presumed to be positively correlated with the quality of life although it appears to be declining as it becomes an obsolete form of communication. The annual rate of real GNP growth is L14. This is a proxy for productivity which is assumed to be welfare enhancing. This position is open to debate and may be reassessed by future philosophers and historians. L15 is % of GNP not for defense expenditures in real terms. It is assumed the less a country is in war, or spends on defense, the more it has to support other welfare needs.

4 Clustering Results

The results of the cluster analysis are summarized in the dendrogram in ...gure 2. A dendrogram (or tree diagram) shows the genealogy of the clusters as they are formed. The distance between clusters is on the horizontal axis and the cluster membership is on the vertical axis.

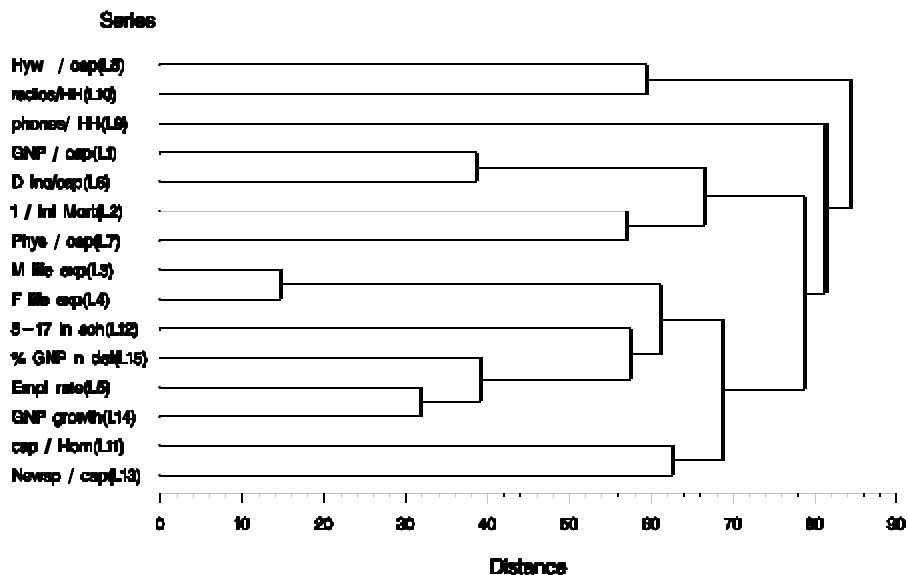


Figure 2: Dendrogram by distances based on $\bar{D}(i; j)$.

This ...gure shows how the clustering method combines the clusters from the case where each is a cluster (15 clusters at the far left hand edge) to the case

in which all the variables are in the same cluster (the far right). For example, from this diagram we can see that the first two series to be combined (the move from 15 to 14 clusters) were the two life expectancy series (L3 and L4). The transition from 14 clusters to 13 was accomplished by the formation of a new cluster containing the employment rate (L5) and the GNP growth rate (L14). This diagram can also be used to investigate the genealogy of any of the clusters to show how they were combined and which series are contained in each.

From Figure 2 we can also see that the distances between most series that are included in the same cluster lie in a range of 15 and 85. Note that the within-cluster distances are relatively small until four clusters are formed. But as we move to combine the series into 3 or fewer clusters we are combining series that are much further apart.

Figure 3 shows the first differences between the distances to form the next cluster. From this plot one can find the cases where the largest changes occur in the distances which produce the next/larger clusters. There is a marked spike in Figure 3 at 3 clusters indicating that a much greater distance was tolerated in order to reduce to 3 clusters than was needed since the reduction to 10 clusters.

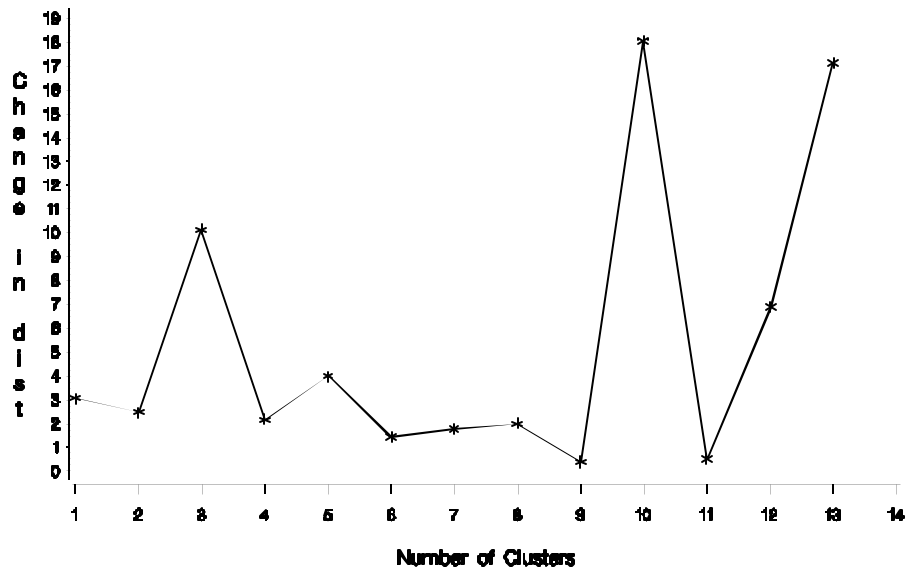


Figure 3: Change in distance to last member added to the cluster.

It would seem that the following three clusters, at least, should be regarded as distinct welfare dimensions: The first cluster contains Highway miles per capita (L8), and the number of radios per household (L10). The second cluster includes GNP per capita (L1), the inverse Infant Mortality Rate (L2), Disposable income per capita (L6), and the number of physicians per capita (L7). In comparing

the distances between the series in this cluster we find that Disposable income per capita (L6) has the smallest average distance to any of the other members of the second cluster. Finally, the third cluster contains male and female life expectancy (L3 and L4), the employment rate(L5), the inverse homicide rate (L11), the % of children aged 5 to 17 in school (L12), the newspaper circulation per capita (L13), the growth rate of GNP (L14) and the % of GNP that is non-defense spending. In the third cluster we find that the growth rate of GNP (L14) is the series that has the smallest average distance to any of the other series in this cluster.

The number of phones per household (L9) is the one series that is still not clustered with any other series at this stage. From table 1 we note that this series is closest to female life expectancy (L4) but it is further from this series than the other members of the third cluster.

Each cluster may be represented by a member closest to its components. The four clusters can then be identified as: 1) the Highway miles per capita (L8)⁹, 2) the disposable income per capita (L6), 3) the growth rate of GNP (L14), and 4) the number of phones per household (L9) cluster. Interestingly, per capita disposable income and the GNP growth rate are statistically distinct dimensions to well-being. The former is often the basis of international inequality analyses, while the latter is the main focus in time series macroeconomics.

4.1 A useful decomposition.

We can decompose the distances into two components. One coming from the residual distributions, the other from the differences in their locations. Consider the hypothetical case where all the series are assumed to have the same innovation distributions except for locations. For example if the series have Normal densities with unit variances, the distance measure $D(i;j)$ can be shown to be given by¹⁰:

$$\bar{D}_n(i;j) = 100 \frac{1}{n} \sum_{t=1}^n \frac{1}{T} \sum_{t=1}^T e^{i \frac{1}{8} (\underline{1}_{it} - \underline{1}_{jt})^2} \text{io};$$

where $\underline{1}_{it}$ is the location of the i th series at time t : Figure 4 is the dendrogram based on $\bar{D}_n(i;j)$. We see that the time profiles that are the most similar in Figure 1 cluster together in this analysis. Note that these clusters are closest to the type of analysis that would be performed in traditional cluster analysis on the basis of distance measures such as the Euclidean. From Figure 4 we note that a number of series are very close together while the relative distances of others is much greater. The cluster that combines the life expectancies, highway miles per capita, radios per household, phones per household, physicians per

⁹The Highway miles per capita is less subject to potential error introduced by the approximation of the radios per household series as detailed in the Appendix.

¹⁰This is due to the result that: $\int_{-1}^1 \frac{1}{\sqrt{2\pi}} \exp(-\frac{(x_i - a)^2}{2}) \int_{-1}^1 \frac{1}{\sqrt{2\pi}} \exp(-\frac{(x_j - b)^2}{2}) \int_{-1}^1 = e^{i \frac{1}{8} (a_i - b)^2}$

capita, GNP per capita, inverse infant mortality, and disposable income per capita is much more homogenous than any other clusters. This is very different from Figure 2.

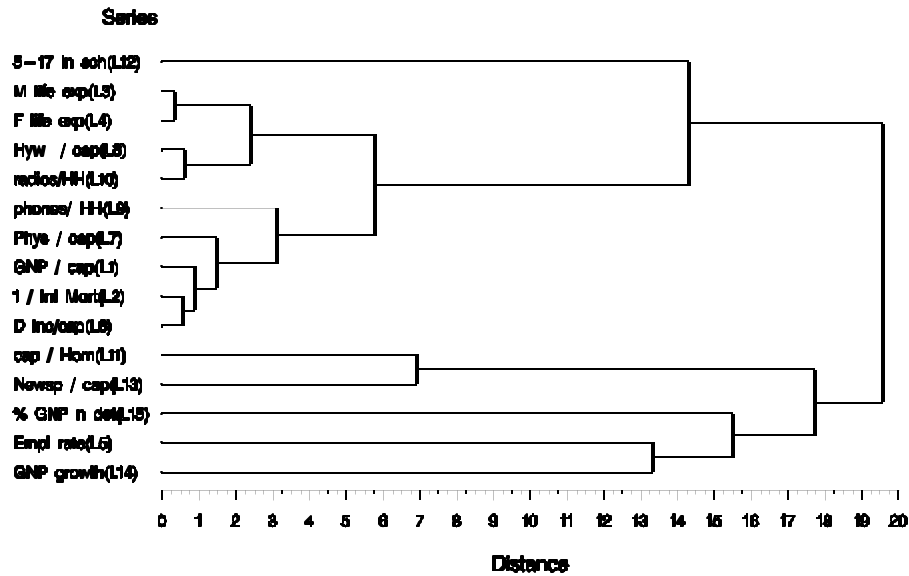


Figure 4: Dendrogram of series using $\bar{D}_n(i;j)$ (under the assumption that each series is normally distributed with a mean of zero and a variance of one).

In order to isolate the effect of the innovation densities on these result, we can cluster only the estimated densities of the innovations. In this case we define a new distance :

$$\bar{D}_r(i;j) = 100 \int_{-1}^1 \int_{-1}^1 h_i(x)^{1/2} h_j(x)^{1/2} dx$$

where the estimated densities $h_i(x)$ and $h_j(x)$ are located at zero and compared directly. These clusters are given by the dendrogram in Figure 5. The most similar densities (inverse infant mortality and the number of radios per household) are clustered together first in this diagram. In addition, we can see that GNP growth (L14), with the most diffuse density estimate, is one of the last series to be included in a cluster. In this way one can evaluate to what extent the clustering using $\bar{D}(i;j)$ is influenced by the nature of the innovations to the series as compared to the distances in the locations.

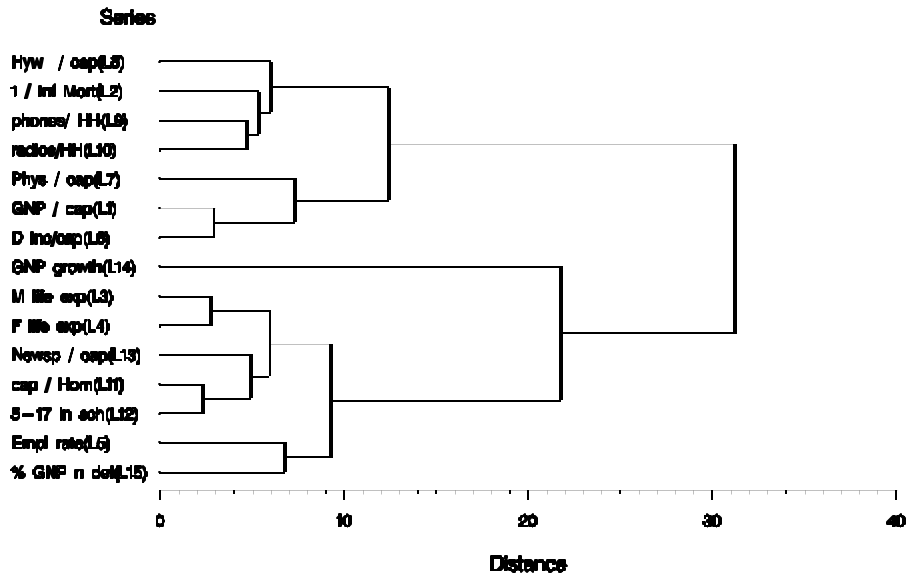


Figure 5: Dendrogram of series using $\bar{D}_r(i;j)$ the part of the distance due to the distances between the distributions of the innovations.

We note that two $I(1)$ variables may or may not be considered as “similar” in the sense of our information measure. Indeed integrated variables are similar in terms of trending properties, and once differenced appropriately, they are $I(0)$ variables which can have very different distributions. It seems perhaps obvious to note that not all stationary variables are similar! The same variables, should they be also cointegrated, produce $I(0)$ variables in linear combinations. This too does not necessarily guarantee “similarity” in the larger sense. Only if the innovation densities are also very similar, or in a largely normal/linear sphere, is cointegration predictive in our sense. There are many series that are known to be cointegrated. But as members of welfare enhancing clusters they would make for very strange bedfellows. Similar confusions are increasingly noted in ...nance with respect to the notion of “random walks”.

5 Conclusions

Our study suggests that there are at least four distinct dimensions in the 15 attribute set we have analyzed. This casts further doubt on unidimensional analyses of well-being. Correlation based studies that entirely focus on linear co-movements between variables may neglect important informational content in the whole distribution of variables.

From the ...rst row of Table 1 one can ...nd the distances between the GNP per capita (L1) and all the other series. Not surprisingly, per capita Disposable income (L6) is the closest. Figure 6 provides a graphic view of the similarity

between the series that best typify each cluster and GNP per capita. Note that, except for cluster 2 which includes GNP per capita, most of the clusters appear to have climbed more slowly in the last 15 years. The most interesting difference between these series is that clusters 1, and 4 seem to be approaching a plateau. And series in cluster 2 appear to be stabilizing. From table 1 we can determine which series is closest to all the other series. The closest on average is the growth rate for GNP (L14), with an average distance of 61.99, followed closely by Male life expectancy (L3) with an average of 62.06 (and much smaller variability). Alternatively by using the minimax criterion we found that Male life expectancy has the smallest maximum distance to any other series in the set.

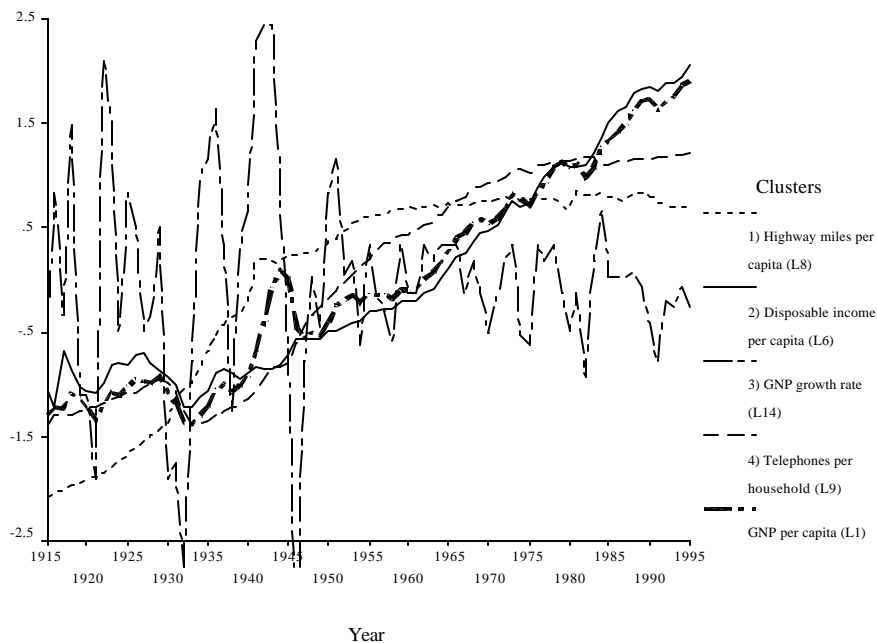


Figure 6: Plots of the series that typify each cluster.

In addition to the identification of clusters, we can aggregate these clusters to create new summary series that incorporate the information contained in the four individual clusters. Using the averaging methods that can be applied here as proposed in Maasoumi (1986) we may create a new combined series. In figure 7 we show three composite indices of wellbeing: a simple average of the 4 representative series, a weighted average based on the number of series in each cluster, and a weighted average based on the inverse of the sum of the square error (SSE) from the ARIMA model fit to the series. This average down-weights those series with large relative variances such as the growth of

GNP which are hard to predict, and increases the contribution of those series with smaller relative variances (are more predictable). It can be seen that the new summary formed by weighting with the inverse of the SSE is closest to the GNP per capita series. However, it also displays the plateau effect.

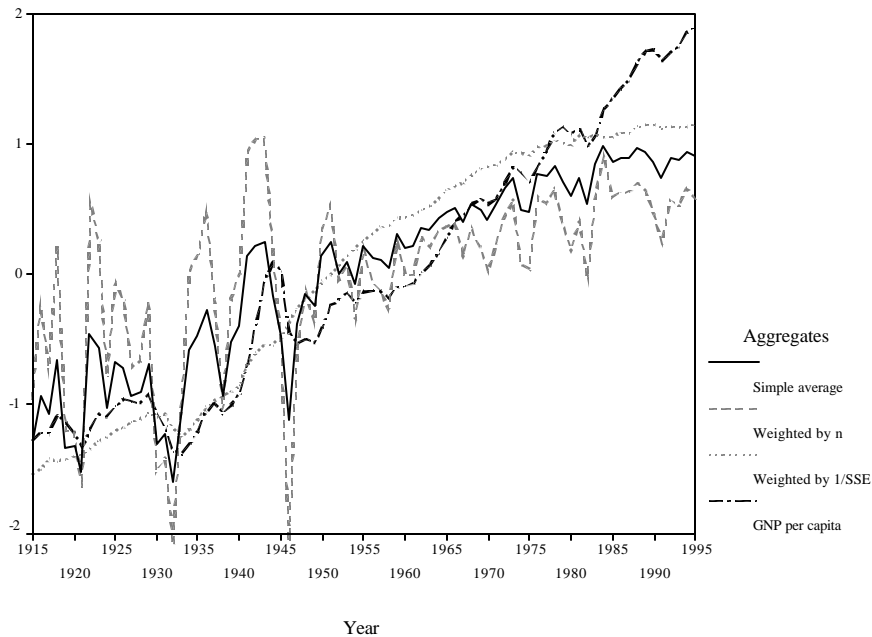


Figure 7: Comparisons of GNP per capita to three alternative weighted averages of the clusters.

Clustering techniques can provide useful non-parametric means of identifying attribute groups that may then be used in multidimensional welfare analyses. Aggregation by clustering may reduce the chance of double counting highly similar attributes. Secondly, aggregation may be desirable if measurement errors and noise in some attribute data is suspected. Thirdly, although the entire time period was used to establish the distance matrix in table 1 (the limit in the summation that defines $\bar{D}(i;j)$ is T), distance matrices for subperiods of the total data series could have been computed to analyze the information content of these series over time.

References

- [1] Atkinson, A. B. and F. Bourguignon (1982), "The Comparison of Multidimensional Distributions of Economic status," *Review of Economic Studies*, 12, 183-201.

- [2] Blomquist, G., M. Berger, and J. Hoehn (1988), "New Estimates of the Quality of Life in Urban Areas," *American Economic Review*, 78, 89-107.
- [3] Bhattacharyya, A. (1943), "On a measure of divergence between two statistical populations defined by their population distributions," *Bulletin Calcutta Mathematical Society*, 35, 99-109.
- [4] Bowman, A. W. (1985), "A Comparative Study of Some Kernel-Based Nonparametric Density Estimators," *Journal of Statistical Computation and Simulation*, 21, 313-327.
- [5] Champernowne, D. G. (1953), "Model of income distribution," *Economic Journal*, 63, 318-51
- [6] Easterly, W. (1997), "Life During Growth," the World Bank (mimeo).
- [7] Epanechnikov, V. A. (1969), "Nonparametric estimation of a multidimensional probability density", *Theory of Probability and its Applications*, 14, 153-158.
- [8] Gibrat, R. (1931), *Les inegalites economiques*, Paris.
- [9] Granger, C. W., E. Maasoumi, and J. Racine (1999), "A Metric Entropy Measure of Dependence," Working paper, Department of Economics, S. M. U., Dallas, TX 75275-0496.
- [10] Gyourko, J. and J. Tracy (1991), "The Structure of Local Public Finance and the Quality of Life," *Journal of Political Economy*, 99, 774-806.
- [11] Hirschberg, J. G., E. Maasoumi and D. J. Slottje, (1991), "Cluster analysis for measuring welfare and quality of life across countries," *Journal of Econometrics*, 50, 131-150.
- [12] Hirschberg, J. G., and J. R. Dayton, (1996) "Detailed patterns of intra-industry trade in processed food," in *Industrial Organization and Trade in the Food Industries*, I M. Sheldon and P. C. Abbott eds., Westview Press, Boulder, Colorado, 141-159.
- [13] Hobijn, B., and P. H. Franses (2000), "Asymptotically perfect and relative convergence of productivity," *Journal of Applied Econometrics*, 15, 59-81.
- [14] Kolm, S-Ch. (1977), "Multidimensional Egalitarianism," *Quarterly Journal of Economics*, 91, 1-13.
- [15] Langley, K. M. (1950), "The distribution of capital in private hands in 1936-38 and 1946-47," *Bulletin of the Oxford University Institute of Statistics*, 12, 339-59.
- [16] Maasoumi, E. (1986), "The Measurement and Decomposition of Multidimensional Inequality," *Econometrica*, 54, 991-997.

- [17] _____(1989), "Composite Indices of Income and other Developmental Indicators: A General Approach," *Research on Economic Inequality*, 1, 269-286.
- [18] _____(1993), "A Compendium on Information Theory in Economics and Econometrics", *Econometric Reviews*, 12, 137-181.
- [19] Maasoumi, E. and J-H. Jeong (1985), "The Trend and the Measurement of World Inequality over Extended Periods of Accounting," *Economics Letters*, 19, 295-301.
- [20] Maasoumi, E. and J. Racine (2001), "Entropy and the Predictability of Stock Market Returns", forthcoming, *Journal of Econometrics*.
- [21] Maharaj, E. A., (1995), "A significance test for classifying ARMA models," *Proceedings of the 1995 Econometrics Conference at Monash, Monash University, Department of Economics*, 219-251.
- [22] Matusita, K. (1967), "On the notion of decision functions," *Annals of the Institute of Statistical Mathematics*, 19, 181-192.
- [23] Mojena, R. (1977) "Hierarchical grouping methods and stopping rules: An evaluation," *Computer Journal*, 20, 359-363.
- [24] McGee, V. E. and W. T. Carlton (1970), "Piecewise Regression," *Journal of the American Statistical Association*, 65, 1109-1124.
- [25] Nordhaus, W. and J. Tobin (1972), "Is Growth Obsolete?," *Fiftieth Anniversary Colloquium V.*, National Bureau of Economic Research, New York: Columbia University Press.
- [26] Piccolo, D., (1990), "A distance measure for classifying ARIMA models," *Journal of Time Series*, 11, 153-164.
- [27] Ram, R. (1984), "Composite Indices of Physical Quality of Life, Basic Needs Fulfillment and Income," *Journal of Development Economics*, 11, 227-247.
- [28] Roback, J. (1980), "The Value of Local Amenities: Theory and Measurement," Ph.D. dissertation, University of Rochester.
- [29] _____(1982), "Wages, Rents, and the Quality of Life," *Journal of Political Economy*, 90, 1257-1278.
- [30] Rosen, S. (1974), "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," *Journal of Political Economy*, 82, 34-55.
- [31] _____(1979), "Wages-based Indexes of Urban Quality of Life," in *Current Issues in Urban Economics*, edited by P. Mieszkowski and M. Straszheim, Baltimore: Johns Hopkins University Press.

- [32] Rutherford, R. S. G. (1955), "Income distributions: a new model," *Econometrica*, 23, 277-94.
- [33] Skaug, H. and D. Tjostheim (1996), "Testing for serial independence using measures of distance between densities", in P.M. Robinson and M. Rosenblatt (eds.), *Athens Conference on Applied Probability and Time Series*, Springer Lecture Notes in Statistics, Springer.
- [34] Sargan, J. D., (1957), "The distribution of wealth," *Econometrica*, 25, 568-90.
- [35] Scott, D. W. (1979), "On optimal and data-based histograms," *Biometrika*, 66, 605-610.
- [36] Sen, A. (1985), *Commodities and Capabilities*, Amsterdam: North-Holland.
- [37] _____(1987), *The Standard of Living*, Cambridge: Cambridge University Press.
- [38] Slesnick, D. (1991), "The Standard of Living in the U.S.," *Review of Income and Wealth*, 37, 363-386.
- [39] Slottje, D. J. (1991), "Measuring the Quality of Life Across Countries," *Review of Economics and Statistics*, 73, 513-519.

Appendix: Data Sources and Notes

The data used in this study are from the following sources:

Economic Report of the President, various editions. Superintendent of Documents, US GPO.

Historical Statistics of the United States: Colonial Times to 1970, Parts 1 and 2, Bicentennial Edition, Bureau of the Census, Dept. of Commerce.

Statistical Abstract of the United States, various editions. Bureau of the Census, Dept. of Commerce.

Notes on particular values used when annual values were missing for the series L6, L7, L9, L10, L12 and L13:

L6 For the disposable income per capita (L6) the years 1912-1916 and 1917-1921 were given as one value so the annual values used here were interpolated.

L7 For the number of physicians used to construct the physicians per capita (L7) series the value for the year before was used for the odd numbered years from 1915 to 1941. For the years 1943-1948, 1971-1974, and 1976-1979 the mean was used. The last value was used for 1956, 1957, 1961, 1981, 1984 and 1988.

L9 The percent of households with phones (L9) were estimated for 1915 to 1919 using the means from later values. From 1982 to 1995 a growth curve estimate is used based on the earlier data and a limiting value of 100%.

L10 The percent of households with radios (L10) were estimated for 1915 to 1921 with a fixed value of .01 then from 1988 on a growth curve estimate based on the earlier data and a limiting value of 100% was used.

L12 The percentage of persons aged 5 to 17 enrolled in school (L12) was linearly interpolated for the odd numbered years from 1917 to 1943. From 1988 to 1993 these values are linearly interpolated from the proximate values.

L13 Newspaper circulations (L13) were estimated for the period from 1915 to 1919 as the mean for the period.