Standardization of longitudinal, aggregate-level data in the Norwegian commune database

Brosveet, Jarle

Jarle Brosveet

Standardization of Longitudinal, Aggregate-Level Data in the Norwegian Commune Database

Introduction

In Norway, communes are the main aggregate-level units for the collection of Bureau of Statistics data. When comparable communal statistics are assembled for different years, data are said to form time-series. The Norwegian Commune Database contains both longitudinal statistics in the form of time-series and data that are incomparable in the sense that they pertain to single years only, depending on what information is available through the Bureau of Statistics.

The Database consists of two parts: 1) the data; 2) the programs that retrieve and manipulate the data. The data part comprises all kinds of official statistics such as data on demography, economy (both tax return and treasury data), industries, schools, elections etc.[1] In principle, the Database will include all published statistics at the commune level that social researchers might want to use. Since some data are more relevant than others, statistics are included according to a priority list. Currently, the Norwegian Social Science Data Services (NSD) has assembled in machine-readable form nearly 8,000 variables from the post-war period and 7,000 variables covering the first 100 years of communal rule from 1838 onwards.

The need for an aggregate-level, longitudinal database such as this one dates back to the late 1950s. The program of Norwegian electoral studies initiated at that time created a demand for various data linked to election statistics. This research program was developed by Henry Valen at the Institute for Social Research in Oslo in cooperation with Stein Rokkan at the Chr. Michelsen Institute in Bergen. Valen concentrated on the most recent data, while Rokkan started work on historical data[2]. In the late 1960s NSD was conceived as a data archiving institution, and the Commune Database was to become one of the main databanks of the Data Services[3].

[1] Rokkan, Stein, and Henrichsen, Bjørn (eds.), Kommunedatabanken. En håndbok for brukere, NSD Handbook No. 2, Bergen 1977.

[2] Valen, Henry, and Rokkan, Stein, The Norwegian Program of Electoral Research, in: Pesonen, Pertti (ed.), Scandinavian Political Studies, Vol. 2, Helsinki, New York 1967.

[3] Rokkan, Stein, and Henrichsen, Bjørn, Building Infrastructures for the Social Sciences: The Norwegian Social Science Data Services, Research in Norway, Oslo 1976.

The activity of the Data Services during the first years of operation concentrated on programming a system for the management of communal data which at the time were divided into separate files of punched cards. One of the tasks of the Data Services was to organize the data so that they could be accessed in an efficient manner. Also, a computer program was designed to keep track of administrative boundary changes at the commune level.

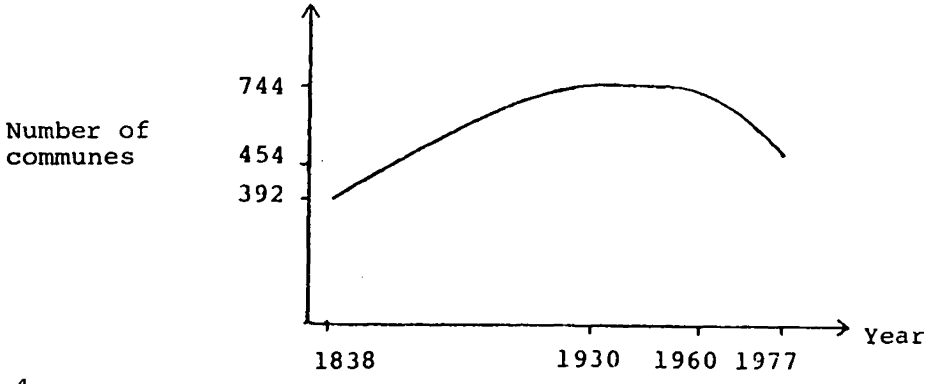
The Commune Database as a programming project has continued up to now. Part of the programs will soon be due for replacement by more efficient database management systems, as the old FORTRAN routines gradually get out of data. However, some features are unique, notably the standardization process which the Database system can carry out.

Boundary Changes

When they were established in 1838, the law enabled communes to divide on their own initiative, which brought about an increasing number of new commune units. After World War II the role of the communes in the nation's economy meant that some kind of *arrondissement* or consolidation had to be instituted. Consequently, during the 1960s more than half the total number of communes were consolidated into larger administrative units[4].

The number of communes, originally 392 in 1838, increased to 744 in 1930. From then on, only small changes took place until the consolidation phase in the 1960s and 1970s. As of writing there are only 454 communes. This development is shown graphically in Figure 1.

Figure 1: Number of Communes 1838–1977



[4] Lie, Suzanne, and Taylor, John G., A Sociological and Geographical Appraisal of Commune Boundary Changes in Norway, mimeo; Oslo 1977.

Figure 2: Nature and Effect of Border Changes

		Effect of Change	
		Creation of New Unit	Extension of Old Unit
Nature of Change	Splits	A	B
	Mergers	C	D

Border changes have the effect of creating new units or extending old ones. Also, the nature of change can be classified as splits or mergers, representing divisions and consolidations respectively. These two dimensions can be grouped in a four-fold table as shown in Figure 2. The table indicates that new units can be established and old ones extended either through split-ups or mergers.

Figure 2 forms a departure point for the coding of border changes. The changes are viewed as having two effects, one for the parting commune and one for the receiving commune. In this way border changes must be recorded bilaterally, and a separate code given to each of the units to indicate whether it is the receiving or parting commune. If more than two communes are involved in the same division or consolidation, a multiple of recorded changes will result in order to register each bilateral relation between parting and receiving communes separately.

The codes used in the Norwegian Commune Database are shown in Figure 3. Two codes are allotted to each of the cells of the table in Figure 2. These codes can be compared to the codes of the Swedish GEOKOD register except for the fact that until the early 1950s, no Swedish commune was extended by split-ups of other established units[5].

Figure 3: Commune Database Border Change Codes

Categories cf. Fig. 2	Code	Explanation
A	1	x is separated from y
A	2	y parts with x
B	3	x incorporates part of y
B	4	y is split to extend x
C	5	x is transformed from y
C	6	y is transformed into x
D	7	x incorporates y
D	8	y is incorporated into x

[5] Öhngren, Bo (ed.), Geokod. En kodlista för den administrativa indelningen i Sverige 1862–1951, Uppsala 1977.

The communal border fluctuations complicate the problem of undertaking longitudinal studies at this level of aggregation. In order to apply frequently used statistical computer programs like SPSS to analyze time-series, data must be standardized to fit the ordinary „flat file" data matrix which presupposes an equal number of variables for all units. The program routines of the Commune Database were designed to do this standardization as part of the retrieval functions invoked when the user requests a particular set of variables.

In principle, there are three ways to standardize a data matrix and take boundary changes into account without resorting to missing value codes indiscriminately to fill up the matrix. The three ways of eliminating troublesome boundary changes are as follows:

(a) units can be dropped
(b) units can be aggregated
(c) properties can be estimated while keeping the actual number of units.

The simplest solution is to drop all units that have been modified. If the drop-out rate proves too high, units can be kept if the alterations involve certain amounts of the resources or significant communes. Still, this is an unsatisfactory solution due to the fact that important values or units might be excluded. If a device can be found to keep as much data as possible in the standardized file, there will be fewer risks of overlooking important variations in the data.

A more agreeable solution is to effect an aggregation of units that merge. The main drawback of this approach is that the process of aggregation will conceal information, making the measures less detailed than in the original file. Also, problems arise when units split and merge simultaneously, in which case a combination of aggregation and omission of units impair the geographical distribution of properties considerably.

The third possibility is to determine the extent to which resources are redistributed as a result of border changes. The fractions indicating resource transfers can be turned into *standardization coefficients* to be applied to data values when time-series analyses are carried out. The idea is that data values for the various units must be standardized as required to facilitate the combination of data from before and after each border change. The standardization process will be most suitable when official statistics on the direct effects of the changes can be found. More imprecise results can be obtained if data have to be estimated or computed. However, some kind of estimation procedure will be requires whenever statistics pertaining to the changes are unavailable.

The last of these three solutions is the most complicated and requires a specially written computer program to be effective. The complications stem from the fact that in the extreme case, separate standardization coefficients will be available for

every property to be redistributed bilaterally. Obviously, the computational procedure can be simplified if more rudimentary coefficients are relied upon. Sometimes, reliance on rudimentary coefficients is required due to the huge amount of work involved in setting up a complete series of standardization coefficients, or coefficients could be partly unavailable due to incomplete statistics.

In Norway, statistics on the direct effects of the border changes are sadly missing. In effect, only two kinds of information are available, namely the size of the area divided or consolidated, and the number of individuals concerned. Size of the area is of little interest in the social sciences, which leaves us with the population figure as our only point of departure, if other measures cannot be estimated.

In the face of the unsatisfactory solution of dropping or aggregating units, locally produced estimates can be relied on to provide coefficients for the standardization process when statistics on the direct effects are unavailable. Theoretically, this strategy constitutes a feasible solution. However, in Norway the idea was impossible to realize in a reasonable way because of the size of the problem. Since 1838 there have been more than 1,000 border changes. Consequently, the number of variables for which standardization coefficients must be found, would be very high. In the end it appeared that we had to rely on a computational procedure based on the available population figures to estimate simplified standardization coefficients.

No doubt a certain amount of error is introduced when a single measure such as the amount of population transferred is used as a redistribution criterion. As always, it is questionable whether resources forming time-series will be correctly redistributed by application of a time-specific coefficient. Also, the population distribution of a particular area has a specific geographic pattern, as people tend to group in agglomerations located close to rivers or fjords. What matters is that the settlement pattern will correlate to a greater or lesser extent with the location of other resources in the same area. Thus the population density can to some extent be used to deduce the geographical distribution of other resources.

Generally, densely populated areas are thought of as being better off in many respects compared to the scattered population of the periphery. To a certain extent measures describing economic well-being and material resources are related to the size and density of the population, whereas the dispersion of natural resources are not. In this way the computational redistribution will sometimes be fairly correct, sometimes wildly misleading, depending on the variables in question.

The bilateral transfer of resources can also be affected by differing resource profiles that characterize the communes. We are now thinking of extremes such as highland vs. lowland, rural vs. urban, inland vs. coastal, etc. Conditions such as these can affect the correctness of the estimated redistribution for instance if one of the communes is predominantly rural showing abundant natural resources, and the other one is predominantly urban, showing a striking amount of welfare facilities.

Also, the population of a commune can be unevenly scattered throughout the area so as to further upset the relationship between people and resources. If a certain percent of the population is ceded when part of one commune is transferred to a neighbouring commune, the transfer can involve the most atypical part of the

population as far as resources are concerned. More often than not transfers affect such atypical groups, as changes take place to improve the homogeneity and internal resource distribution of the communes.

Traditionally, the distribution of population and resources are unequal in Norwegian communes, so many of the adverse comments that we have made, will apply[6]. In spite of these apparently unfavourable conditions, the computational approach was selected for implementation. It remains to be seen how this approach affects the data, as no analysis has been carried out yet to evaluate the effects on time-series analysis. However, for the time being the simple computational approach seems best, as the dropping of units is deemed more erroneous, aggregation is partly prevented by concurrent split-ups and consolidations, and the summoning of local expertise to produce individual coefficients is too expensive and unreliable to be of much use.

The Standardization Process

Even if careful estimates are made for every border change, some resources will be difficult to redistribute in a sensible way. Naturally, a crude computational procedure will to some extent give wrong results unless precautions are taken. The problems of the described heterogeneous distribution of people and resources cannot be fully resolved, but some of the undesirable effects can be removed.

Clearly, buildings and institutions cannot split. Neither can measures such as per capita income and public expenditure. Generally, when there are few members in a class of objects, or if the resource has no obvious relation to the population distribution, values cannot split between parting and receiving communes by being computed. Conversely, resources which appear to have a more continuous or homogeneous character throughout the population can be redistributed according to the amount of population transferred. Demographic variables lend themselves most favourably to such treatment and can be redistributed by means of the population figure almost unconditionally.

Unfortunately, it appeared that most of the variables in the Commune Database belong to the unsplittable category and had to be assigned missing values for both parting and receiving communes. To decrease the amount of missing values, *computability codes* were introduced. A computability code is associated with each variable to tell the computer program what to do when a pair of communes is subject to modifications.

[6] Lie and Taylor, op. cit.

The computability codes used in the Commune Database to indicate the computational measures to be taken, are as follows:

(a) proportional computation
(b) computation for consolidations only
(c) transfer if values are identical numerically
(d) no computation or transfer

Proportional computation means that the percent of total population transferred will be the coefficients used by the program to redistribute the variable values for the communes in question. Computation in the case of consolidations means that a summation of values will take place only for border changes belonging to categories C and D in Figure 2. These are cases of 100 percent transfer of resources because the parting unit ceases to exist. Computation only if the values of the parting and receiving communes are identical in numerical terms means that both parting and receiving communes retain their values unchanged. No computation means that missing data will be filled in for both parting and receiving communes, as no computation, transfer or summation of values can take place.

Each variable in the Database is assigned only one computability code upon judgement of which code is most suitable throughout the country. No local conditions that represent exceptions will be considered by the computer program during processing. Neither are the cumulative effects of successive changes taken into account. In this way there will be no check on the admittedly infrequent cases when a split neutralizes a previous merger or vice versa, as the system has no way of knowing if part of a commune has been affected twice.

The change codes explained in Figure 3 are only used for setting up the standardization coefficients. During actual processing these coefficients plus the computability codes constitute the constants relied on by the program to produce the standardized matrix. Figure 4 will give you an idea of the information that the program reads to enable the standardization process. First, the program will read the commune identification numbers and the information on years of establishment and discontinuation as indicated in Figure 4a. This information helps the program determine the actual number of communes for any single year. Next, information on border changes are accessed as specified in Figure 4b.

The data part of the Database is stored in a socalled transposed matrix (i. e., variablewise instead of the ordinary casewise structure) using the exact number of communes that existed in the year of data collection. This information is available to the program as part of the variable label. By means of the establishment file the program will identify which data elements belong to which communes, as the ordering of the data elements corresponds to the sequencial order of the commune identification numbers.

Next, the program will read which year the user has specified as the year of standardization, as well as the list of variables that the user wants to extract from the Database. By means of the border change file and the computability codes the program will now standardize the data by applying the coefficients year by year from the year of data collection till the year of standardization. When all variables

Figure 4a: File of Establishment and Discontinuation

Commune identification number	Year of discontinuation	Year of establishment (default is 1838)
0101		
0102		
0103		
0104		
0111		
0112	1964	
.		
.		
.		
0402	1964	1965
.		
.		
.		
1201		1972
.		
.		
2030		

Figure 4b: File of Border Changes

Parting commune	Year	No. of receiving$_1$	Coeffi-cient$_1$	Receiving commune$_1$	Coeffi-cient$_2$	Receiving commune$_2$
1264	1945	1	28.335	1265		
0218	1947	1	100.000	0301		
1041	1948	1	1.542	1003		
0814	1949	1	2.661	0802		
.						
.						
.						
0537	1962	2	8.383	0538	91.616	0536
.						
.						
.						
1941	1974	1	31.884	1942		
1855	1974	1	100.000	1805		

have been treated in this manner, the user will have the data transposed back into the rectangular flat file structure ready for analysis by SPSS or any other statistical program package[7].

Further Capabilities of the Database

The Commune Database was originally thought of as a bank of post-war data from Bureau of Statistics publications and magnetic tapes. The supporting computer routines were meant to offer a solution of the standardization problem for the post-war period only, so they were not written with a general purpose in mind.

The most serious restriction of the original routines was the ability to standardize only forward in time. This solution was chosen to minimize the various errors that can be inflicted by the computational process. Most significantly, the post-war period brought about a majority of consolidations, as indicated in Figure 1. The consolidations resulted in an increasing number of larger and more resource rich communes as we approach the 1970s. Standardizing backwards in time would split these resource rich units into several relatively impoverished communes on false assumptions, as the heterogeneous character of the communes before consolidation cannot be deduced. The extreme case happens when a city has been merged with neighbouring communes that are predominantly rural.

Things become more complicated as the Database is extended back to 1838. Now the user must be able to choose whether standardization should be effected forwards or backwards in time. If all data belong to the period after 1930, standardization forwards will give the least errors, as there is a majority of consolidations in this period. Data belonging entirely to the period before 1960 should be standardized backwards because of the divisions[8]. This is a useful rule of thumb which can be trusted in most cases. If the user wants to combine pre-1930 with post-1960 data, no advice can be given. In this case a close scrutiny of modifications and variables is required to see how standardization affects the data.

Because of the extended time-span of the present Database, some additional requirements must be satisfied to have an acceptable retrieval system for time-series data of nearly 150 years. One of the most important improvements is a report facility that users can involve to get a modification report on paper. This modification report will consist of a listing in tabular form of the variable values before and after

[7] Brosveet, Jarle, Teknisk dokumentasjon for Kommunedatabanken 1945–1977, NSD Report No. 15, Bergen 1977.

[8] Aarebrot, Frank H., and Kristiansen, Bjarne, Norwegian Ecological Data 1868–1903, NSD Report No. 9, Bergen 1976.

standardization. By consulting the list, the user can inspect changes to see if corrections are needed.

Users also want to control other parts of the standardization process, such as the coefficients and the computability codes to be used, as the decisions made at the time of implementation will not suit all needs. Though the user is discouraged from meddling with the computational process, errors can be corrected more efficiently by specifying an alternative set of coefficients and codes instead of updating values in the standardized file.

A seemingly dispensable improvement is the selection feature that enables the user to have a subset of standardized communes on file. However, it has proved important that the standardized file produced by the Database is as ready for analysis as possible. Even if selection routines are found in most statistical program packages, users find it tedious to start an analysis run by filtering out units they never wanted from the Database.

Another requirement is the need to add data temporarily. This requirement stems from the wish of users to include data that are not part of the Database. In this way the user can have his own data standardized and included on the same file as any Commune Database variables that he retrieves. User-supplied data may be derivates or raw data from sources other than the Bureau of Statistics that NSD does not include in the regular updating of the Database. Naturally, if user-added data are of sufficient general interest, they will be included permanently, if no clauses are attached.

Future Development

The Commune Database is but one of the conceivable systems for the retrieval of locality data. A host of other units at various levels also need attention. Separately, NSD has made available a Census Tract Database for 1960 and 1970 data aggregated from person level records. To some extent the Census Tract Database takes care of the linkage to the Commune Database, though its primary objective is to standardize data at the census tract level[9].

It is hardly feasible to build separate systems for each locality level, so some kind of general solution is called for. No doubt, experience from the Commune Database project and the Census Tract Database project will offer some guidance on how to design a generalized linkage system for aggregate data. In particular, a

[9] Alvheim, Atle, NSDs Kretsdatabank. Datainnhold og brukerveiledning, NSD Report No. 18, Bergen 1977.

general linkage system will be needed because of the difficulties of drawing definite conclusions based on data for only one level of aggregation.

Another major development trend is the setting up of a Cartographic Database that will enable data from the Commune Database to be drawn as maps[10]. Since the geographical distribution of variable values often appears to form striking patterns, maps will be the best way of discovering regional differences. Statistical graphics can enhance the display of geographic patterns even more, so a graphics system coupled to the map-drawing routines is being tested[11].

The utility of the Database can be improved even further if data were available at the individual level. The 19th century censuses will be available soon as they can be released without violating personal integrity as protected by the Person Register Law. Modern data are harder to get hold of, as most personal and locality identifications are likely to be removed from the data when they are made available to the Data Services. However, academically conducted surveys and background information on public figures do not suffer from this restriction, although such data are mostly too specific to be of general interest.

The Norwegian Social Science Data Services will also store specific interest data as required, but as far as the Commune Database is concerned, most potent research opportunities will result when census data are combined with communal time-series. NSD will take steps to improve the linkage of such data as one of our major projects for the 1980s.

[10] Sande, Terje, Cartographic Database for Time-Variable Data, EPD 25 (December 1977), a contributed paper to the European Meeting on Regional Databases for Computer Cartography, Bergen 1977.

[11] Alvheim, Atle, Figur: A Program for Point Graphics, EPD 25 (December 1977), a contributed paper to the European Meeting on Regional Databases for Computer Cartography, Bergen 1977.