A socio-economic classification of small areas in the county of Östergötland

Linköping may 2008

Lars Walter
Elin Eriksson
A socio-economic classification of small areas in the county of Östergötland

Report 2008:2

Linköping may 2008

Lars Walter
Elin Eriksson
Foreword

The Centre for public health sciences is a department in the County council of Östergötland, Sweden.

This report provides background information and references for a socio-economic classification of the county of Östergötland. In addition to analysing the interaction between two SES levels, area-based SES by itself also serves as a useful tool for analysing and describing variations in factors of interest in the population. This is not a paper intended for a scientific publication.
## Contents

**BACKGROUND** ............................................................................................................................... 5

**MATERIALS AND METHODS** ........................................................................................................... 6

  SMALL AREAS ................................................................................................................................. 6

  DATA .................................................................................................................................................. 6

  *Statistics Sweden* .......................................................................................................................... 6

  *The County Council of Östergötland* ......................................................................................... 7

**THE 2006 POPULATION BASED SURVEY IN ÖSTERGÖTLAND** ..................................................... 8

**METHODS** ...................................................................................................................................... 8

**RESULTS** ........................................................................................................................................ 10

  *Description of the final classification* ....................................................................................... 13

  *Evaluation of the results by health and lifestyle factors* ............................................................ 13

  *Geographical presentation of SES* .............................................................................................. 14

**DISCUSSION** ................................................................................................................................... 15

**CONCLUSION** ................................................................................................................................. 15

**REFERENCES** .................................................................................................................................. 16

**APPENDIX 1** .................................................................................................................................... 17

**APPENDIX 2** .................................................................................................................................... 19
Background

There is a well-known and well-documented relationship between socio-economic status (SES), and health and lifestyle factors, such as smoking habits and physical activity [1–4]. The discussion involves two levels of socio-economic status for a person: one individual and the other area-based [4]. The individual level refers to a person’s individual financial resources; the area-based level concerns the socio-economic status of the neighbourhood where the person is living. Having knowledge about the two levels makes it possible to analyse the interaction between individual and area-based SES [5]. For example, is the health of an individual with low SES living in a deprived area different from the health of an individual with low SES living in a non-deprived area?

This report focuses on area-based SES in Östergötland. In addition to analysing the interaction between the two SES levels, area-based SES by itself also serves as a useful tool for analysing and describing variations in factors of interest in the population, for example:

- What is the relationship between health and area-based SES?
- What is the relationship between lifestyle factors and area-based SES?
- Which health care centres cover the most deprived areas?
- Is there an unequal distribution of socio-economy in Östergötland?

Östergötland was classified into five different levels of SES in 1999 [6] and then updated in 2002 [7]. The classifications were performed by the Centre for Public Health Sciences, using data mainly from 1994 in the first classification, and data from 1999 for the update. The multivariate technique cluster analysis was used to find areas with similar SES, based on variables describing income, financial assistance and education.

This report describes the results of an updated SES analysis based on data from 2004–2006. The aim has been twofold: to update the SES and the geographical boundaries of the small areas.
Materials and methods

Small areas

We use key-code areas (NYKO\textsuperscript{1}) as small areas for this application. Key-code areas are geographical areas that constitute the smallest administrative areas in each municipality. We used the so-called 4-digit level, which in urban areas means that each NYKO comprises several blocks with similar characteristics. There are 942 key-code areas in the county of Östergötland.

Data

Statistics Sweden

Statistics Sweden (SCB) is a central government authority for official statistics and other government statistics. They have the responsibility for coordinating and supporting the Swedish system for official statistics. SCB also keeps digitalised maps of all key-code areas in Sweden. The personal linkage between different registries is a unique 10-digit personal identification number. All data were aggregated in key-code areas by SCB. To protect personal integrity, SCB gives full information only if the aggregated data is based on at least 3 individuals. We used the following variables, updated for 2004–2006, from SCB:

- **Population** covers persons aged 0 years or more by age and sex. The variable used in this report is the proportion of retired persons, defined as persons aged 65 years or more.

- **Immigration** covers the number of persons born outside of Sweden aged 0 years or more. The variable used in this report is the proportion of immigrants.

- **Unemployment** covers the number of unemployed persons aged 18–64 years. The variable used in this report is the proportion of unemployed persons aged 18–64 years.

- **Education** covers the number of persons aged 16–64 years attending compulsory school, upper secondary school or university as the highest level of education. The variables used in this report are high education (defined as the proportion of persons, aged 16–64 years, with university as the highest level of education) and low education (defined as the proportion of persons, aged 16–64 years, with compulsory school as the highest level of education).

- **Disposable income.** Apart from salary, income consists of capital income and various types of allowances, after deduction of all taxes, measured in Swedish kronor (SEK) for persons aged 16 years or more. Three income

---

\textsuperscript{1} In swedish: Nyckelkodsområden
variables, based on persons aged 16 years or more including persons with no income, are used in this report: mean disposable income, high income and low income. High and low income are defined as the proportion of persons having income above the 75th percentile and below the 25th percentile of disposable income, respectively.

- **Social assistance.** The variable used in this report is mean social assistance among families (members aged 20 years or more), in Swedish kronor (SEK).

- **Housing supplement.** The variable used in this report is mean housing supplement among persons aged 65 years or more, in Swedish kronor (SEK).

- **Accommodation** covers the number of families aged 20 years or more living in detached houses or blocks of flats. The variable used in this report is called building characteristics. An area is classified as a detached house area if the proportion of families (members aged 20 years or more) living in detached houses is above 0.50, otherwise it is classified as a block of flats area.

- **Number of days reporting sick** is measured as days paid from the Swedish Social Insurance Agency, among persons aged 16–64 years, and includes among other things sickness benefits, disability pensions and temporary disability pensions. The employers pay the first 14 sick days. The variable used in this report is the mean number of days paid from the Swedish Social Insurance Agency among persons aged 16–64 years.

**The County Council of Östergötland**

The County Council in Östergötland keeps registries of all visits to institutional and non-institutional care. The linkage between key-code areas and individuals is done by the geographic coordinates of the residence for each individual. All data were aggregated in the five final SES levels.

We collected four variables from the County Council registries:

- **Caries.** The variable used in this report is the proportion of persons aged 6 years without any experience of caries in 2006, measured by deft (due to caries, decayed, extracted or filled primary teeth).

- **Psychiatric illness.** The variable used in this report is the proportion of persons with main diagnoses of ICD F00–F99 from institutional care in 2006.

- **Heart disease.** The variable used in this report is the proportion of persons with main diagnoses of ICD I20–I25 from institutional care in 2006.

- **Visits to a general practitioner (GP).** The variable used in this report is the mean number of visits to a GP at health care centres in 2006.
The 2006 population based survey in Östergötland

A population-based survey was performed in Östergötland in 2006. The questionnaire was sent to 13 440 persons aged 18–84 years. The survey was administrated by SCB and the response rate was 54 percent. The questionnaire contained among other things questions about self-rated health and lifestyle factors. Key-code areas were linked to each respondent from the SCB registries and the results were aggregated in the five final SES levels.

We used the following variables from the survey:

- Alcohol consumption. The estimated proportion of persons, aged 18–84 years, with risky drinking behaviour. Risky drinking is defined as 14 glasses per week for men and 9 glasses per week for women. One glass is a proxy for 12 grams of alcohol.

- Overweight. The estimated proportion of persons, aged 18–84 years, who are overweight, defined as a body mass index of 25 or more.

- Physical activity. The estimated proportion of persons, aged 18–84 years, who are not regularly physical active.\(^2\)

- Tobacco use. The estimated proportion of persons, aged 18–84 years, who are daily smokers.

- Self-rated health. The proportion of persons, aged 18–84 years, reporting excellent or very good health in question 1 of the SF36 Health Survey [8].

Methods

The previous analyses, in 1999 and 2002, were based on cluster techniques. The aim was to cluster key-code areas into homogenous groups based on variables related to SES.

One advantage of cluster techniques is that they aggregate observations, i.e. key-code areas, combining information from several variables with practically no statistical assumptions. On the other hand, it is not obvious when to stop aggregating and obtain the final number of groups. This must more or less rely on suitable subjective reasoning. To find the number of groups, the previously analyses used Ward’s hierarchical clustering method. The analysis supported five clusters, which is also suitable from a practical point of view. Thereafter the K-means method was used to assign each observation to one of the five clusters [9]. All variables were standardised by subtracting the mean and dividing by the standard deviation.

\(^2\) A person who is physically active for at least 30 minutes every day is defined as regularly physical active.
The choice of variables to include in the cluster analysis is also more or less subjective. To achieve the final ranking of groups into SES levels, each group of key-code areas is described by variables included in the analysis as well as variables not included in the analysis. Finally, the distribution of key-code areas into different SES levels needs to be acceptable. The two extremes (the best SES level and the worst SES level) should comprise approximately 5–15 percent of the key-code areas and each of the three intermediary levels should comprise more than 15 percent. This will create a socio-economic gradient that will point out the really deprived as well as the most privileged areas.

If we assume that the number of clusters is predefined, the evaluation of the final results involves examining two criteria:

- **First criterion.** Socio-economic indicators should show a reasonable trend along the SES levels. Evaluation of this criterion is done by ranking the clusters into SES levels and describing them based on available information.

- **Second criterion.** The distribution of key-code areas among the SES levels should show a reasonable shape. Evaluation of this criterion is done by ranking the clusters into SES levels and describing the distribution of key-code areas along the SES gradient.

If some of the socio-economic indicators do not agree with the socio-economic ranking, the clustering should start all over again with a new set of variables, still assuming that the number of clusters is predefined. In addition, if clustering produces a distribution of key-code areas that does not follow the preferable distribution, the procedure has to be done again with a new set of variables.

The present analysis aims to redo the cluster analysis from 2002 based on updated data and, if possible, to find a simpler model. Compared to a model based on cluster analysis, which involves more than one variable, a one-variable model would be simpler in several ways. The single-variable model would constitute a simple index and the first criterion above is used to find the best variable. There is no need for the second criterion because we can decide the cut off points in the index resulting in a preferable distribution.

We have used five levels of SES for practical reasons. It is the same number of levels that have been used in previous analyses in Östergötland, for example on dental status among children and heart disease (6,7).
Results

Key-code areas comprising less than 10 inhabitants were removed from the analysis together with key-code areas with extreme values for one or more variables. The remaining 882 key-code areas that were included in the analysis are described in Table 1.

Table 1. Description of \( n=882 \) key-code areas included in the analysis (percent of inhabitants)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Median</th>
<th>Standard deviation</th>
<th>25th/75th percentiles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retired</td>
<td>18</td>
<td>16</td>
<td>13</td>
<td>11/22</td>
</tr>
<tr>
<td>Immigrants</td>
<td>7</td>
<td>5</td>
<td>7</td>
<td>3/9</td>
</tr>
<tr>
<td>Unemployed</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>0/5</td>
</tr>
<tr>
<td>Low educated</td>
<td>18</td>
<td>17</td>
<td>9</td>
<td>12/23</td>
</tr>
<tr>
<td>Low income</td>
<td>26</td>
<td>24</td>
<td>11</td>
<td>19/31</td>
</tr>
<tr>
<td>Social assistance</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>0/3</td>
</tr>
<tr>
<td>Proportion of detached areas</td>
<td>73</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Three different models were analysed. The first model is based on the same variables as the analysis from 2002, i.e. income, economic assistance and education. The second model is based on variables that produced the “best” result according to the two criteria mentioned earlier. In both models, key-code areas were aggregated by the K-means technique. The third model is constructed as an index based on a single variable. To evaluate the results according to the first criterion, a graph is presented for each model. The graph shows the relationship between the socio-economic gradient and seven variables that are assumed to correlate with socio-economy. The ranking of the five clusters according to SES is made subject to producing as good trends as possible in the graphs.

The first model involves four variables:
- mean income
- mean social assistance
- mean housing supplement
- proportion with high education.

The results for model 1 are presented in Figure 1a. Some variables do not follow the preferable non-decreasing trend along the SES levels; see for example low education and immigrants. In addition, the important SES indicator, low income, shows a poor trend, at least at one point along the SES axis.
Figure 1a. Results from model 1.

The second model involves:
- proportion with high income
- mean housing supplement
- building characteristics
- proportion unemployed

and is the “best” model based on the cluster technique. The results for model two are seen in Figure 1b and show the same drawbacks as model 1. For example, low income, retired and immigrants show poor trends along the SES gradient.

Figure 1b. Results from model 2.
Table 2 shows the distribution of key-code areas between the SES. The preferable shape of the distribution is as symmetric as possible with approximately 5–15 percent at each extreme. The two models show neither a symmetric distribution nor the preferable percentage at the two extremes.

**Table 2. Distribution of key-code areas between the SES (percent of areas)**

<table>
<thead>
<tr>
<th>Socio-economic status</th>
<th>Highest</th>
<th>Lowest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>24</td>
<td>45</td>
</tr>
<tr>
<td>Model 2</td>
<td>28</td>
<td>22</td>
</tr>
<tr>
<td>Model 3</td>
<td>10</td>
<td>20</td>
</tr>
</tbody>
</table>

Figure 1c presents the results based on the third model. This model involves only a single variable that acts as a continuous index, allowing the distribution of the key-code areas into the five SES levels as follows: 10 percent, 20 percent, 40 percent, 20 percent and 10 percent, in order from low to high SES. According to that distribution, several variables were examined and the best result was achieved with disposable income above the 75th percentile as the index variable.

Comparing the graphs of the three models (Figure 1a–c) shows that model 3 does not have the same drawbacks as models 1 and 2. This means that model 3 performed better than models 1 and 2 according to the first criterion above. The second criterion will always be fulfilled with a single variable index.
**Description of the final classification**

Table 3 and Appendix 1 describe the five SES levels based on different demographic and socio-economic variables. Table 3 is based on the number of inhabitants at each SES level and Appendix 1 is based on key-code areas. Table 3 is interpreted as follows: for example, the proportion of immigrants among the inhabitants in the highest SES is 6 percent and 25 percent in the lowest SES. Appendix 1 describes the distribution of the key-code areas at the different SES levels with respect to different variables. For example, the 10th, 50th and 90th percentiles of the 88 key-code areas at the highest SES level are 0, 5 and 10 percent respectively, according to the proportion of immigrants.

**Table 3. Description of final classification (percent of inhabitants)**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Socio-economic status</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Highest</td>
</tr>
<tr>
<td>Number of key-code areas</td>
<td>88</td>
</tr>
<tr>
<td>Number of persons</td>
<td>31291</td>
</tr>
<tr>
<td>Women</td>
<td>50</td>
</tr>
<tr>
<td>Children (0–14 years)</td>
<td>13</td>
</tr>
<tr>
<td>Adolescents (15–29 years)</td>
<td>12</td>
</tr>
<tr>
<td>Aged 65–80 years</td>
<td>3</td>
</tr>
<tr>
<td>Aged 81 years or more</td>
<td>6</td>
</tr>
<tr>
<td>Immigrants</td>
<td>2</td>
</tr>
<tr>
<td>Unemployed</td>
<td>9</td>
</tr>
<tr>
<td>Low education</td>
<td>4</td>
</tr>
<tr>
<td>Low income</td>
<td>15</td>
</tr>
<tr>
<td>Social assistance</td>
<td>0</td>
</tr>
</tbody>
</table>

**Evaluation of the results by health and lifestyle factors**

Some health-related register variables were used after the analysis. These variables are used to show the relationship between health and SES (see Table 4). The gradient of days reporting sick, non-caries and visits to GP along the SES levels is very clear, with worse health at the lowest SES level. Psychiatric illness and heart disease are also most frequent at the lowest SES level, but the gradient is not that clear.
Table 4. Health related register variables (mean per inhabitant and percent of inhabitants)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Socio-economic status</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Highest</td>
</tr>
<tr>
<td>Mean number of days reporting sick</td>
<td>27</td>
</tr>
<tr>
<td>Non-caries (percent)</td>
<td>84</td>
</tr>
<tr>
<td>Psychiatric illness (percent)</td>
<td>5</td>
</tr>
<tr>
<td>Heart disease (percent)</td>
<td>2</td>
</tr>
<tr>
<td>Mean number of visits to GP</td>
<td>0.89</td>
</tr>
</tbody>
</table>

Some lifestyle variables from the 2006 population survey are presented in Table 5. It shows a relationship between lifestyle factors and socio-economic status. Areas with the lowest SES have the highest proportion of daily smokers, persons who are overweight and physically inactive. More people living in privileged areas report excellent or very good health. The relationship between risky consumption of alcohol and SES is not that obvious.

Table 5. Lifestyle factors and self-rated health (percent of inhabitants)

<table>
<thead>
<tr>
<th></th>
<th>Socio-economic status</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Highest</td>
</tr>
<tr>
<td>Daily smokers</td>
<td>11</td>
</tr>
<tr>
<td>Risky consumption of alcohol</td>
<td>12</td>
</tr>
<tr>
<td>Not regularly physically active</td>
<td>73</td>
</tr>
<tr>
<td>Overweight</td>
<td>44</td>
</tr>
<tr>
<td>Excellent/very good health</td>
<td>54</td>
</tr>
</tbody>
</table>

Geographical presentation of SES

In Appendix 2, we present maps with SES levels for the main population centres in each municipality. Each dot on the map represents one residence, which means that each dot represents more than one inhabitant. Key-code areas at the highest SES level are dark green and areas at the lowest SES level are yellow. The geographic boundaries of the key-code areas are based on digitalised maps kept by Statistics Sweden. One municipality, Valdemarsvik, has no updated map and has therefore been omitted from Appendix 2.
Discussion

The main advantage of the cluster technique in this application is that it can handle information from several variables simultaneously. On the other hand, it also introduces some difficulties in the analysis:

- It is difficult to explain the final model because the structure of the model cannot be explicitly expressed.

- The distribution of key-code areas into different clusters is not controllable. This means that the variable in the model is dependent on the distribution and vice versa.

- The order of the cluster along an SES gradient is not obvious. This means that the investigator has to rank the cluster according to several variables simultaneously, which can be a difficult task.

The two latter points create some problems when it comes to updating the analysis. Updating normally aims to reclassify the key-code areas based on updated values for the variables in the model. What if the distribution of the key-code areas does not follow the most favourable distribution? The worst case scenario is that the analysis has to start all over again to find new variables that will fulfil the two criteria used to judge the performance of the analysis.

What drawbacks are associated with the simpler model using a single variable as an index? The main disadvantage is that a single variable may not contain enough information for an acceptable classification. One way to find out if that is the case is to look at the information contained in the graphs in Figure 1a–c. The graphs of the present analysis reveal that a single variable shows at least as good, if not better, performance, compared to the two cluster analyses.

The benefits of a single-variable model are obvious if we look at the three points above. The model is easy to explain and it is up to the investigator to decide on the distribution. The third point is trivial. Moreover, a single-variable model is flexible in both the number of SES levels and how to distribute the key-code areas along the levels. Finally, updating a single-variable index is straightforward.

Conclusion

A single-variable index can be used to perform an overall classification of key-code areas into five SES levels. The proportion of persons with disposal income above the 75th percentile should be the index variable to classify key-code areas into five SES levels with the following distribution from highest to lowest: 10 percent, 20 percent, 40 percent, 20 percent and 10 percent.
References


## Appendix 1

### Description of the final classification based on key-code areas

<table>
<thead>
<tr>
<th>Variable</th>
<th>High SES</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Low SES</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10%</td>
<td>50%</td>
<td>Mean</td>
<td>90%</td>
<td>10%</td>
<td>50%</td>
<td>Mean</td>
<td>90%</td>
<td>10%</td>
</tr>
<tr>
<td>Immigrants</td>
<td>0.00</td>
<td>0.05</td>
<td>0.05</td>
<td>0.10</td>
<td>0.00</td>
<td>0.05</td>
<td>0.06</td>
<td>0.11</td>
<td>0.00</td>
</tr>
<tr>
<td>Low education</td>
<td>0.00</td>
<td>0.08</td>
<td>0.09</td>
<td>0.19</td>
<td>0.00</td>
<td>0.14</td>
<td>0.14</td>
<td>0.20</td>
<td>0.10</td>
</tr>
<tr>
<td>Social assistance</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.03</td>
<td>0.00</td>
</tr>
<tr>
<td>Low income</td>
<td>0.00</td>
<td>0.15</td>
<td>0.16</td>
<td>0.27</td>
<td>0.00</td>
<td>0.19</td>
<td>0.21</td>
<td>0.29</td>
<td>0.16</td>
</tr>
<tr>
<td>Unemployed</td>
<td>0.00</td>
<td>0.01</td>
<td>0.02</td>
<td>0.04</td>
<td>0.14</td>
<td>0.03</td>
<td>0.03</td>
<td>0.05</td>
<td>0.00</td>
</tr>
<tr>
<td>Children (0–14 years)</td>
<td>0.00</td>
<td>0.21</td>
<td>0.20</td>
<td>0.30</td>
<td>0.00</td>
<td>0.20</td>
<td>0.19</td>
<td>0.26</td>
<td>0.08</td>
</tr>
<tr>
<td>Adolescents (15–29 years)</td>
<td>0.00</td>
<td>0.12</td>
<td>0.13</td>
<td>0.21</td>
<td>0.12</td>
<td>0.15</td>
<td>0.16</td>
<td>0.22</td>
<td>0.09</td>
</tr>
<tr>
<td>65–80 years</td>
<td>0.00</td>
<td>0.08</td>
<td>0.09</td>
<td>0.21</td>
<td>0.09</td>
<td>0.10</td>
<td>0.10</td>
<td>0.18</td>
<td>0.06</td>
</tr>
<tr>
<td>81 years or more</td>
<td>0.00</td>
<td>0.00</td>
<td>0.02</td>
<td>0.05</td>
<td>0.03</td>
<td>0.02</td>
<td>0.03</td>
<td>0.07</td>
<td>0.00</td>
</tr>
<tr>
<td>Detached areas</td>
<td>0.86</td>
<td></td>
<td>0.91</td>
<td></td>
<td>0.81</td>
<td></td>
<td>0.59</td>
<td></td>
<td>0.18</td>
</tr>
</tbody>
</table>
Appendix 2