Small Area Unemployment Statistics System

Calculating labour force data also for small areas

MultiRacio Ltd.

Contents

1. Small Area Unemployment Data

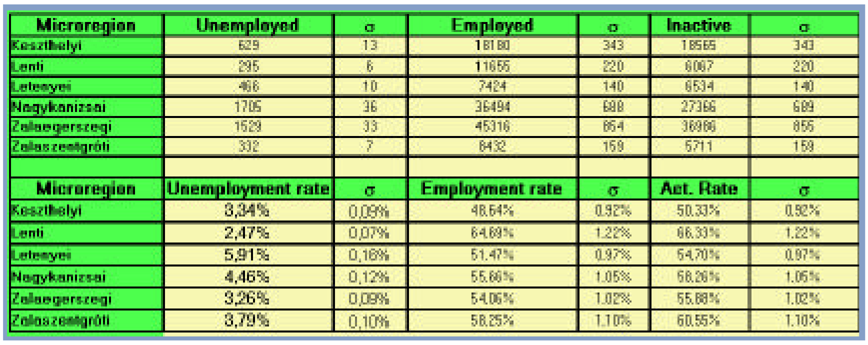
For the operation and investigation of a modern market economy it is inevitable to obtain data describing the labour market: economically active population, employment and unemployment, activity and unemployment rates and other indicators.

These data can only be used correctly if they fulfill the following conditions:

- Not only the number of the registered unemployed is important, but also the more realistic employment and unemployment numbers **defined by ILO (International Labour Organization)**. These data may come from different surveys (usually household statistics, or labour force surveys) and statistical calculations.

- In addition to country level numbers, labour data of smaller areas like counties, statistical regions, or individual settlements need also to be calculated. These data are needed for the investigation of the differences of the labour market in different areas of the country, which is important when distributing development resources. Another important aspect is the possibility of obtaining European Union resources aimed to regional development. A growing proportion of EU resources can be attained by economically intergrated smaller areas, even if they are in different administrative units or even in different countries. Labour force data for these areas can be calculated only if we have a sound system of small area statistics.

Table 1. Labour force data of statistical microregions and the error (σ) of the estimated data.



Microregion	Unemployed	σ	Employed	σ	Inactive	σ
Keszthelyi	629	13	18180	343	18565	343
Lenti	295	6	11655	220	6067	220
Letenyei	466	10	7424	140	6534	140
Nagykanizsai	1705	36	36494	688	27366	689
Zalaegerszegi	1529	33	45316	854	36986	855
Zalaszentgróti	332	7	8432	159	5711	159
Microregion	**Unemployment rate**	**σ**	**Employment rate**	**σ**	**Act. Rate**	**σ**
Keszthelyi	3,34%	0,09%	48,64%	0,92%	50,33%	0,92%
Lenti	2,47%	0,07%	64,69%	1,22%	66,33%	1,22%
Letenyei	5,91%	0,16%	51,47%	0,97%	54,70%	0,97%
Nagykanizsai	4,46%	0,12%	55,66%	1,05%	58,26%	1,05%
Zalaegerszegi	3,26%	0,09%	54,06%	1,02%	55,88%	1,02%
Zalaszentgróti	3,79%	0,10%	58,25%	1,10%	60,55%	1,10%

2. The SAUS System

Labour force data for small areas in Hungary are provided by the Small Area Unemployment Statistics System (SAUS), originally developed and maintained by MultiRáció Ltd. SAUS is accepted by the Hungarian government as official data source since 1999.

The system combines data from labour force surveys with data from administrative sources using spatial and time series statistical methods. This procedure filters out most of the sampling error of the survey data, and yields more reliable and spatially distributable estimations than the direct estimation method.

Application of this method also saves the considerable cost of getting data with the same reliability in the traditional way of increasing sample size. Calculations showed that the annual cost of a survey with increased sample size would be an order of magnitude higher than the total annual cost of the SAUS system including operation, development and maintanance.

The system is built in a way that it can be easily modified to handle data of other countries or other types of surveys.

Main features of the system:

- Gives up-to-date data about unemployment, emloyment and other labour force indexes in Hungary, on regional, county and small area (statistical microregion and labour market area) level.

- Data can be used for international comparison and analysis, as they are prepared following the international standards of ILO.

- The system applies optimized calculation methods, that can lower the statistical errors of the estimations to an acceptable level also for small areas.

- In addition to estimating small area data at a given date, it also provides a time series view, applying state-of-the-art time series analysis methods.

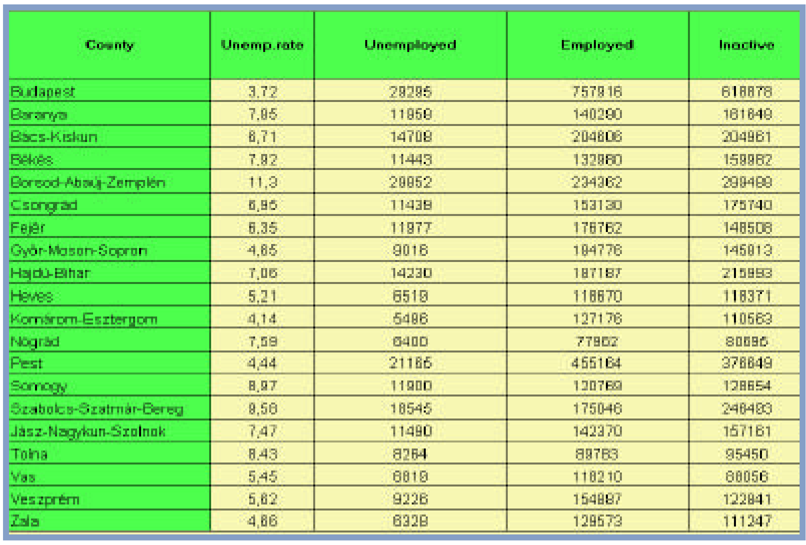
- The system is built in a modular and easy-to-handle way. It can be installed and operated without any specific knowledge, and does not require any special hardware. The easily customizable reports can be presented in various formats, or serve as a starting point of further analyses.

- The system includes a website built on the same database, which can present the results in tabular and graphical form on the web.

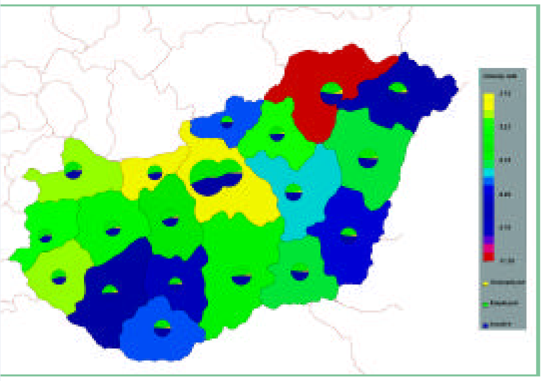
The estimations have been prepared for the National Employment Office since 1999 as official data. These data are used by various government organizations for developing economic strategies.

You can see below some examples of the tables and graphs provided by the SAUS system.

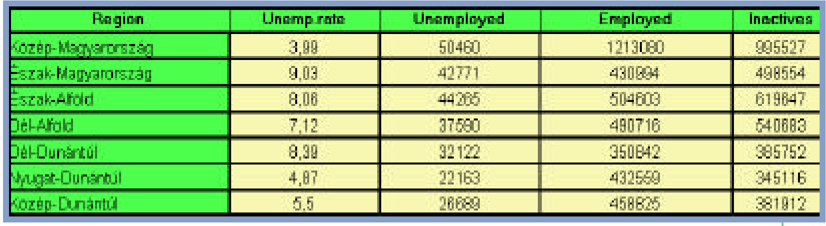
1. Counties of Hungary, January 2003



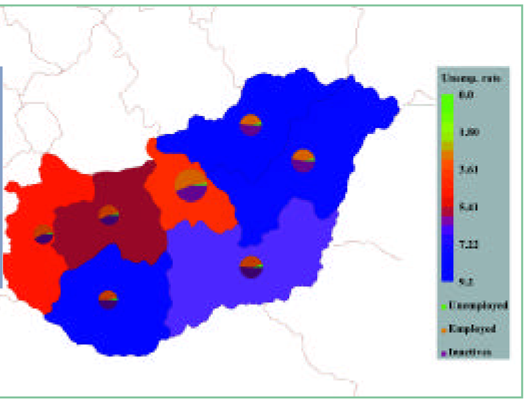
County	Unemp.rate	Unemployed	Employed	Inactive
Budapest	3,72	29295	757916	618878
Baranya	7,85	11959	140290	181649
Bács-Kiskun	6,71	14708	204806	204961
Békés	7,92	11443	132980	159982
Borsod-Abaúj-Zemplén	11,3	29852	234362	299499
Csongrád	6,95	11438	153130	175740
Fejér	6,35	11977	176762	148508
Győr-Moson-Sopron	4,65	9016	184776	145913
Hajdú-Bihar	7,06	14230	187187	215993
Heves	5,21	6519	118670	118371
Komárom-Esztergom	4,14	5486	127176	110563
Nógrád	7,59	6400	77962	80696
Pest	4,44	21186	455164	378649
Somogy	8,97	11900	120769	128654
Szabolcs-Szatmár-Bereg	9,56	18545	175046	248483
Jász-Nagykun-Szolnok	7,47	11490	142370	157161
Tolna	8,43	8264	89763	95450
Vas	5,45	6819	118210	88056
Veszprém	5,62	9226	154887	122841
Zala	4,86	6329	129573	111247

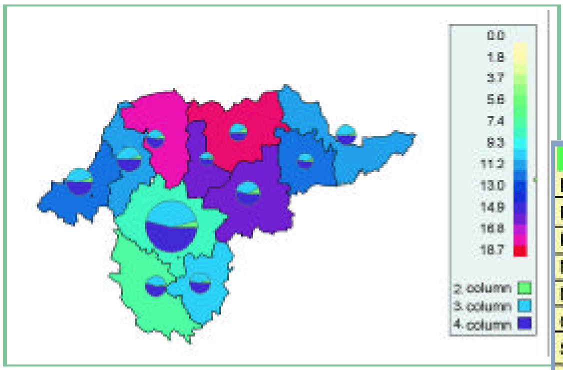


2. Statistical regions of Hungary, January 2003

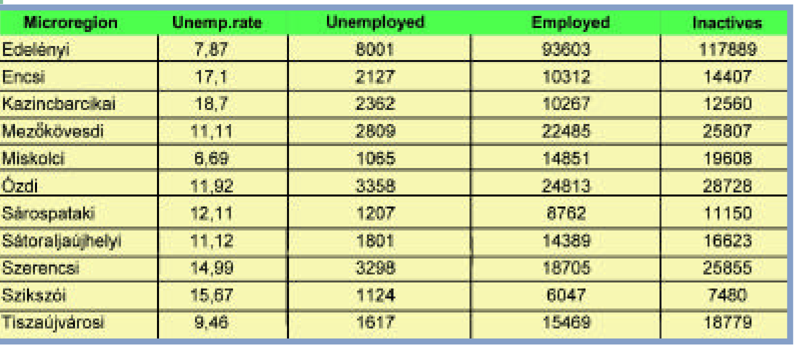


Region	Unemp.rate	Unemployed	Employed	Inactives
Közép-Magyarország	3,99	50460	1213080	995527
Észak-Magyarország	9,03	42771	430994	498554
Észak-Alföld	8,06	44265	504603	619647
Dél-Alföld	7,12	37590	480716	540883
Dél-Dunántúl	8,39	32122	350842	395752
Nyugat-Dunántúl	4,87	22163	432559	345116
Közép-Dunántúl	5,5	26689	458825	381912



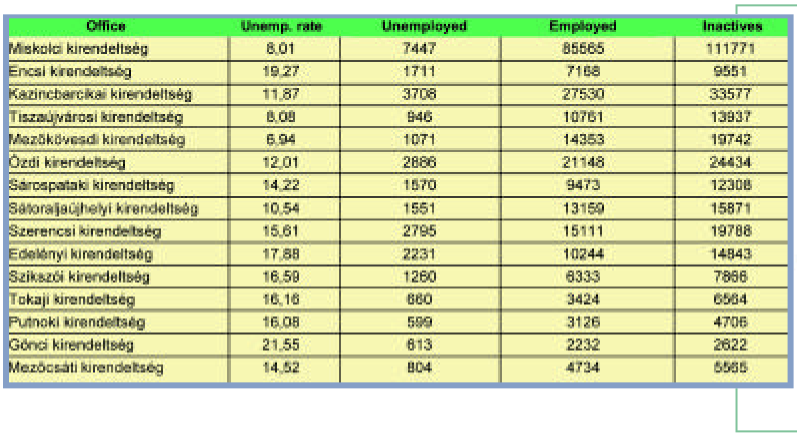


3. Statistical microregions of Borsod-Abaúj-Zemplén county, January 2003

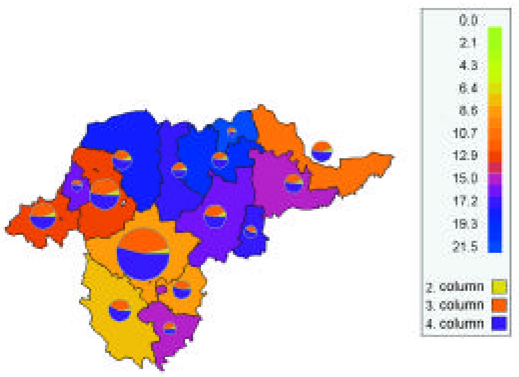


Microregion	Unemp.rate	Unemployed	Employed	Inactives
Edelényi	7,87	8001	93603	117889
Encsi	17,1	2127	10312	14407
Kazincbarcikai	18,7	2362	10267	12560
Mezőkövesdi	11,11	2809	22485	25807
Miskolci	6,69	1065	14851	19608
Ózdi	11,92	3358	24813	28728
Sárospataki	12,11	1207	8762	11150
Sátoraljaújhelyi	11,12	1801	14389	16623
Szerencsi	14,99	3298	18705	25855
Szikszói	15,67	1124	6047	7480
Tiszaújvárosi	9,46	1617	15469	18779

4. Labour offices in Borsod-Abaúj-Zemplén county, January 2003



Office	Unemp. rate	Unemployed	Employed	Inactives
Miskolci kirendeltség	8,01	7447	85565	111771
Encsi kirendeltség	19,27	1711	7168	9551
Kazincbarcikai kirendeltség	11,87	3708	27530	33577
Tiszaújvárosi kirendeltség	8,08	946	10761	13937
Mezőkövesdi kirendeltség	6,94	1071	14353	19742
Ózdi kirendeltség	12,01	2886	21148	24434
Sárospataki kirendeltség	14,22	1570	9473	12308
Sátoraljaújhelyi kirendeltség	10,54	1551	13159	15871
Szerencsi kirendeltség	15,61	2795	15111	19788
Edelényi kirendeltség	17,88	2231	10244	14843
Szikszói kirendeltség	16,59	1260	6333	7866
Tokaji kirendeltség	16,16	660	3424	6564
Putnoki kirendeltség	16,08	599	3126	4706
Gönci kirendeltség	21,55	613	2232	2622
Mezőcsáti kirendeltség	14,52	804	4734	5565



3. Methodology

Input data

In a modern economy, there are usually two independent sources of unemployment data:

- one built on the registration of unemployed people, and
- another one coming from houshold surveys measureing unemployment and employment data according to ILO definitions.

Both sources have their considerable advantages and serious disadvantages as well. Big advantage of the registration data is that they are up-to-date and can be got in any spatial details. Their disadvantage is that they can not be used in international comparisions, and the calculation of the unemployment rate needs other data that is not detailed in the same way (economically active population). Household surveys, on the other hand, can measure unemployment based on an international definition and standard method, but cannot provide detailed and quick information, becasue of the relatively small sample size. Combination of these two data sources, together with some supplementary data (e.g. population, monthly and quarterly labour statistics) provides the best solution, keeping the advantages and avoiding the disadvantages of both sources.

ILO standards define the following labour statistics terms:

Employed: every person aged 15-74 years who worked at least 1 hour for pay in the given period (the so called reference week, which contains the 12. day of the month), or was absent from work only temporarily (because of sickness, vacation, bad weather,...).

Unemployed: every person aged 15-74 years who

- did not work during the reference week,
- was actively looking for work during the four weeks previous to the interview,
- was available to start work within the two weeks following the survey week or were waiting to start a new job within a period of 30 days.

Economically actives: every person aged 15-74, who belongs to either of the above groups, that is the employed and the unemployed together.

Economically inactives: All other persons in the 15-74 age group. A special group of these is the passive unemployed, who would like to work, but they did not seek work because they do not see any hope to find a suitable job.

For the best estimations of the above data and the corresponding indexes the SAUS system builds on the following input data sources:

- Official mothly report about the number of registered unemployed.
- Quarterly report about the results of the labour Force Survey (LFS) performed by the Hungarian Central Statistical Office (HCSO).
- Annual population data

Estimator functions and calculation of statistical error

The aim of our procedure is to prepare reliable monthly and quarterly estimations of the number of employed and unemployed people in the regions, counties and small areas, using the best combinations of various data sources, estimaton and time series analysis methods.

The direct estimator of the survey, originally published by the Central Statistical Office was without bias, but had an unacceptably high variance when applied to smaller areas due to the relatively small sample size. It is especially true in the case of monthly data, where it is inevitable to use time series analysis methods. Fluctuations in quarterly data are of course much smaller, so the time series methods are not absolutely necessary.

In case of counties, at a given point in time, we used the following (generalized) regression modell with a regression coefficient fitted to country wide data:

$$M^{*}=Y^{*}+B_{0}(X-X^{*})$$

where M^* is the estimated unemployment in the county, X is the number of registered unemployed (Employment Office data), Y^* is the direct estimator given by CSO, X^* is the estimation of the registered unemployment coming also from the CSO's labour Force Survey, and B_0 is the country wide linear regression coefficient between the previous two quantities.

Our procedure is the corrected synthetic estimator method. The resulting county and country level values ($X_{\text{county}}(t)$) are then adjusted to the country level direct CSO estimator (${}^{d}X_{\text{country}}(t)$), taking into account the requirement of additivity:

$${}^{\text{corr}}X_{\text{county}}(t)=X_{\text{county}}(t)*{}^{d}X_{\text{country}}(t)/\Sigma_{\text{county}}X_{\text{county}}(t)$$

The standard error of the estimation (empirical variance) is calculated by the so called „jackknife" method. Due to the pecularities of the sampling method, this procedure can only be applied on the county level, and is described here in short:

We form n J_i subsamples of the whole sample of the CSO labour force survey. A subsample, or „jacknife" sub-county is the sample without the i-th primary sampling unit (PSU). An additive statistics on the county can be calculated by linear interpolation based on its „jacknife sub-county" value. If we perform this calculation for all sub-counties, we can use the resulting (i=1...n) „pseudo-values" for the variance and standard error calculation.

Regional level estimations can be got by simply adding the corresponding county level results, as the current statistical regions in Hungary consist of complete counties.

Annual estimations are prepared using the same methods, just the base is the sum of the quarterly or monthly data.

Small area estimations („distribution")

Estimations are prepared with the above method only for county or larger areas. These can be regarded as „small areas" in international comparisions, and these procedures give considerably better results than the previously applied unreliable direct method. However, the main goal was to get good data also for areas smaller than the counties. County level numbers are therefore distributed to smaller areas. The small areas of main interest in Hungary are:

- Employment Office (EO) districts, which are the administrative units of unemployment care.
- Statistical microregions used in the CSO reports when publishing geografically distributed data.

The two types of areas are approximatelly of the same order of magnitude, counties consist of about 6-12 of these. They are overlapping in most cases, but there are considerable differences as well. The SAUS calculates data for both types of districts independently, but the method of distribution is the same.

Small area estimation is done using the population-demand method, based on the county level results. The following formula calculates the number of employed at a given point in time:

$$F_{i}=\frac{F_{i}^{n}}{\sum_{i}\left(F_{i}^{n}\frac{N_{i}}{N_{i}^{n}}\right)}\frac{N_{i}}{N_{i}^{n}}F \quad ,$$

where all values denoted by N refer to the population and the F values to the number of the employed people in the age group between 15-74 years. Upper index n denotes the last census data, lower index i refers to the small area, and the F without indices shows the county wide employment. Distribution of the county unemployment is done simply according to the proportions of the registered unemployment numbers in the corresponding small area.

Time series analysis (Kalman-filter)

All the methods mentioned above work with isochronous data, that is they are based on correcting the survey results based on other data of the same point of time. The time series methods get the additional information from data of the past. The combination of the two types of methods gives an optimum use of all the information we have for improving the reliability of our estimations.

Our system uses the "structural time series model", which has been successfully applied by the US Buro of Labour Statistics for processing labour market survey results. The other widely used approach, the ARIMA modelling can be treated as part of the more general structural model.

The model

If we compare the time series of unemployment based on the monthly survey and the series based on the registry of the National Labour Centre, we immediately notice the difference: the survey data series jump up and down, while the registry data change much smoother. This is due to the so called sampling error which deviates the data measured on a sample (e.g. on 1000 interviewed person) from the real population value. Our model should therefore account for this feature: we assume that the measured values are the sum of the real population value and a random sampling error:

$$y(t) = \Gamma(t) + \eta(t) .$$

The sampling error is assumed to behave as random noise, but still having some regular pattern due to the sample selection method: only one third of the sample members are changed each month. This behaviour is modelled by an ARMA (short for autoregressive-moving average) process, multiplied by a term which accounts for the changing variance of the sampling error:

$$\eta(t) = \gamma(t) e^*(t) .$$

The population value is structured further: We assume that it is the sum of a regression, a seasonal and a trend component.

$$\Gamma(t)=X(t)\beta(t)+T(t)+S(t).$$

where $X(t)$ is the explanatory variable, $\beta(t)$ is a stochastically varying regression coefficient.

The *regression component* relates the survey data to the registry data. It describes how the real number of unemployed depends on the number of those who registered at the employment service. This relation is not constant over time, and this is why we need the additional seasonal and trend components.

The *seasonal component* $S(t)$ tries to capture the short term changes in the composition of unemployed labour force, e.g. young job-seekers leaving school at summer time or the seasonality of temporary jobs in agriculture.

The *trend component* $T(t)$ represents long term changes in the behaviour of the labour force, usually only a small correction factor to the components described above.

All these model components have several parameters, which are usually unknown. These parameters specify the individual model in a more general class of models. A procedure of model fitting should be performed to find an estimated value for them. After the initial model fitting is done, the model can be used to get estimates or forecasts of the values measured in the CSO survey.

The Kalman-filter

Estimation and forecast is done by the Kalman-filter algorithm, which is a general and optimal method to analyse and estimate structured time series. Iterating through the time series from its beginning to the end and then back to its origin, the filter collects information on the behaviour of the observed series and at the same time, prepares estimates of the underlying structure of the data. This structure is represented by the decomposition of the measured value to the model components defined in the above section. This is how we obtain a time series of the estimated sampling error, trend, seasonal and regression components. All we need to do is to subtract this sampling error from the measured value, and we have the smoothed estimates of the real population value.

Forecast is done in a similar way. In the first step, the filter is run on the existing series, then it can be used to model the behaviour of the series in later points of time.

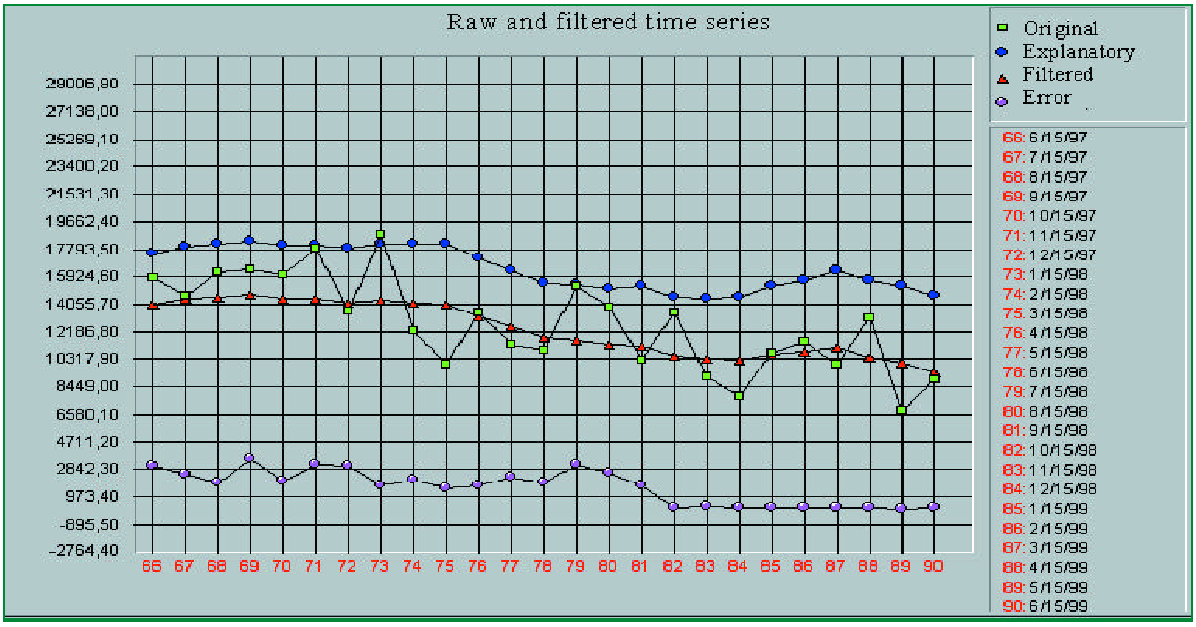
One of the advantages of the Kalman-filter algorithm is that it not only gives estimates of the value itself, but estimates of its variance. Thus we also get the reliability of our estimations or forecasts.

Model fitting

The model fitting procedure uses the same input data as the data estimation procedure, and is performed as an iteration process with the goal of maximizing the likelihood function. There are several methods for this, all of them known as Maximum Likelihood methods. We currently use the EM-algorithm also used by the BLS in the United States. There seems to be a need, however, to look for a quicker algorithm, which can be combined with EM to increase the performance of the estimation procedure.

Example

As an illustration of the type of results we get with the Kalman filter, we included the following chart:



The Input time series is the number of unemployed in the county Zala, estimated from the CSO monthly survey. We applied the model described above, using the number of registered unemployed as an explanatory variable of the regression part of the model. This Explanatory time series is also plotted and it shows evidently less variation than the more noisy survey time series. The Filtered time series is the actual result of Kalman filtering. It includes the regression, seasonal and the trend components. The Error time series shows the expected error of the Kalman filter estimation process. It is surprisingly small, but it must be noted that it does not include the possible error of the model selection.

Time series adjustment (benchmarking)

Once a year, when all data of the preceding year are available, the SAUS system performs an adjustment of the data of the previous year. This is necessary to make the estimations conform to the HCSO annual county data.

Time series adjustment is a procedure of fitting two time series coming from different sources.

If we have two different time series of the same variable, but with different sampling frequency, the two series give usually different annual values, as the direct annual value is usually based on a larger sample, and is more accurate than the monthly estimations.

Time series adjustment („benchmarking”) provides an optimal way of fitting the less accurate monthly data to the more reliable annual value. Optimal way means that we reach conformance by the smallest possible distortion of the in-year behaviour of the monthly series.

One of the most widespread benchmarking procedures is the Denton method, which belongs to the class of constrained quadratic optimization methods.

Represent the monthly data by vector

$$z=[z_1,z_2,\ldots z_n]$$

and the more relieable annual series by vector

$$y=[y_1,y_2,\ldots y_n].$$

We look for a vector

$$x=[x_1,x_2,\ldots x_n]$$

which

a) minimizes the deviance from the original vector z by a constrain function (this is the sum of squares of the first differences, in the case of the Denton method)

b) fulfils the condition that the annual sums of the monthly series are equal to the corresponding values of the annaul series.

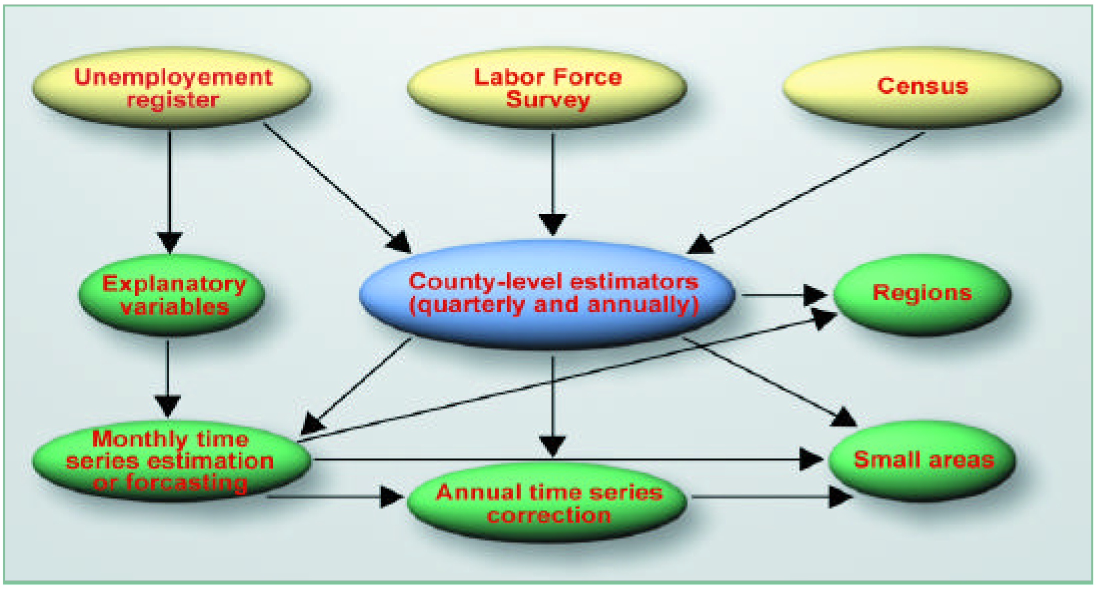
Thus the minimum should be reached by the constraint condition where k is the number of the in-year periods (12, if we have monthly series), and m is the number of the years.

$$\sum_{t=(T-1)*k+1}^{T*k} x_t = y_t, \quad T=1,2,\ldots m$$

If we solve this mathematical problem (minimization of a quadratic function with a constraint condition) by the Lagrange mutiplicator method, we can get the vector x as a linear matrix transformation of y and z.

The Denton method does not give error values for the adjusted time series. If it will be required, the system can be changed to the more elabourate Cholette-Dagum method, which also calculates error estimates

Flowchart of SAUS


Unemployement register
Labor Force Survey
Census
Explanatory variables
County-level estimators (quarterly and annually)
Regions
Monthly time series estimation or forcasting
Small areas
Annual time series correction

4. The IT Solution

During the actual development of the SAUS system by MultiRáció we followed the following principles:

Reliability — the SAUS system is running under Linux operation system, data are stored in a relational database, and regular automatic backups help to avoid data loss.

Modularity — the system consists of individual software modules, which are only connected through the database. They get their input and place their output in the database. The modules can be tested and developed independently.

Maintainability — we used only mainstream technologies, avoided solutions that would require special knowledge.

The SAUS syetm requieres the following **software environment:**

- Relational database engine MySQL 3.22 or newer.
- Interpreter of the Ox matrix programming language and its SsfPack library.
- X Window to run the graphical user interface program.
- OpenOffice compatible office software for the reports and especially EuroOffice to prepare mapped graphs.

Database

MySQL server running under Linux. This combination is a widespread solution in the practice of these days, and because it is the leading open source technology, its reliability is better than similar products on the market.

Graphical User Interface

Operating the system is made easy by a graphical user interface program. This program can be used to access all functions of the system, but for the routine tasks we can also use simple scripts to automate the actions.

Data conversion, explanatory variables, estimator functions

We developed three individual programs for data input and conversion, for the calculation of the explanatory variables (magy) and for the estimator functions and their variance (bf13).

Kalman-filter

The Kalman-filter algorithm (time series estimation and forcasting) is programmed using the Ox programming language and the SsfPack library.

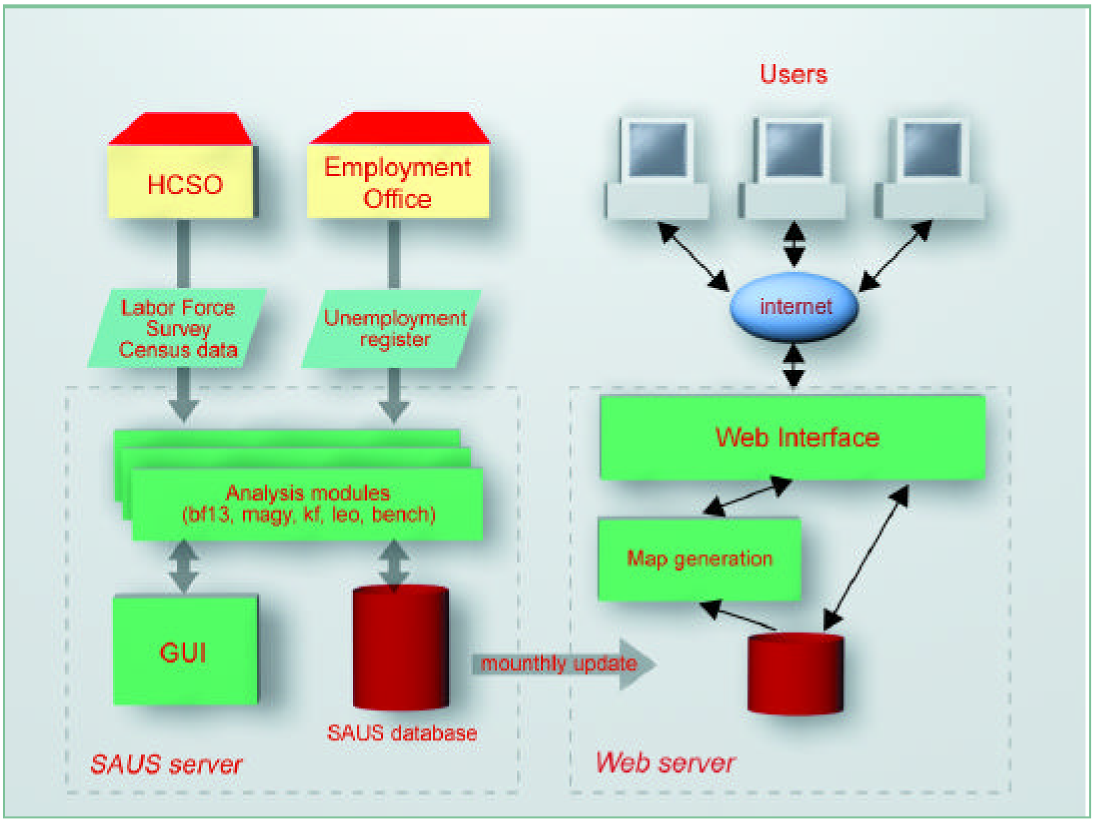
Benchmarking, small area estimation

Time series adjustment („benchmarking") and the distribution of county data to small areas is also done by independent Linux programs.

Reports

Reports are prepared by EuroOffice templates. EuroOffice is a full featured office productivity suite based on the open standard OpenOffice.org. The tables of the reports can be viewed and used for further analysis by any of the office programs in the OpenOffice family, e.g. StarOffice, or OpenOffice.org itself. The template has a direct database connection, which makes it easy to update the tables and the graphs. The reports can be saved in MS Excel format as well. EuroOffice also supports mapped charts, which can visualize country or county wide spatial distributon of unemployment data.

Block diagram of the SAUS system



Web site

The small area labour force data can be found at the web site of the National Employment Office: http://www.afsz.hu