Dr. Udo Maack

SCORUS Session 3: **Results of recent regional and urban analysis**

Organiser : Berthold FELDMANN (Eurostat)
Chairperson : Berthold FELDMANN
Discussant : Dev VIRDEE

Title of paper :	The effect of changing regional classifications on statistical indicators
Name of author :	Dr. Udo MAACK

F1: Good afternoon ladies and gentlemen.

I'm presenting the result of my experience in Turkey, where we have implemented a new regional database which is published in the internet.

The regional database of TURKSTAT, the national statistical office of Turkey was build up and is maintained by a small group of 4 persons. They are responsible not only for the maintenance but also have to create special reports like "Portray's" of selected regions. They don't collect and prepare statistical data, they are provided by the domain oriented production units, but nevertheless so the resources are very limited, the procedure must be very effective.

To enhance the workflow of uploading data, we are just developing a module to provide functions for a semi automatic procedure. During the analysis of this procedure, we detected some requirements which are the kernel of my presentation today. Results look very self-evident and should be available everywhere, but implementations in statistical offices are very few.

F2: What happens to statistical data, if the regional classification changes

I have structured my presentation into three sections

First: The role of regional indicators, why and for what purpose we collect and prepare regional indicators

Second: How can the regional classification change and what happens in statistics

Third: - and this is the most important part – what we have to provide to handle this changes.

- How we have to document the changes and
- How we have to recalculate our data.

F3: The role and characteristics of regional statistical indicators

In all statistics we like to describe a phenomenon, that's unexceptionable.

To describe the development of phenomena in a region, we build time series over the statistical values.

Each region is described by several characteristics of that region. This means Regional statistics, as a multi domain approach, has to compile data from many domains. The data from different domains are linked together, using the regional identifier. To do this without, each individual domain has to use the same regional classification.

Using the same territorial division is one of the preconditions to make regions comparable. Two comparisons are of interest:

The comparisons of regions
Within the country and
Between regions internationally.

To compare regions of different size, the values must be normalized. Area or population are the prevailed factors most used.

Statistical data mostly serves to measure the result of political decisions, the success or the failure. The political audience generally is oriented on actual division of the territory and the development, not taking in account the division of the past.
A changing territorial division which is not taken in consideration in the statistical reports, reduces the validity and herewith the plausibility and the credibility of statistics.

F4: How can the regional classification change and how is this recognized in statistical data?

As we can see in this example, where we have a table showing some datum per end of the year of Turkish provinces between 1995 and 2004, some values are missed e.g. for "Düzce", as marked in red / yellow.

This is results from creating a new province in 1999 and was statistically in effect since January 1st 2000.

Because the total area of Turkey didn't change during that period, the new province was created taking parts from other existing provinces.

This development is not taken into account for the data shown for 1995 to 1999. The ascertained data are still assigned to the regions classified at that time.

The portion of counted values in the regions which belong to the moved areas have to be reassigned, indices which are based on area or population have to be adjusted accordingly.

Before going deeper, let us look

F5: Which changes may occur?

The diagram shows - as an example - the development of the number of provinces in Turkey, which rose in five steps from 73 in 1990 to 81 in 2000.

But not only new regions are relevant for our question, also the change of borderlines between provinces, e.g. resulting from the reassignments of districts or municipalities to adjoining provinces.

How often this happened and how relevant it is, shows the next foil.

F6: Sixty eight changes on province level happened during the period 1990 to 1999. I call it internal changes, shown in blue, the pinck part of the columns represent the new provinces. In total there are 76 changes which require an adjustment of the statistical values.

The effects in detail I have to skip, regarding the limited time, but if someone likes to ask questions, please don't hesitate to ask me.

How these changes can be handled in an NSO, I like to present now. I have headed this chapter with

F7: Provisions to avoid a false reporting.

There are three actions which have to be taken:

Establish a central maintained server containing the regional classification over time.

Provide mechanism and procedures to adjust or to recalculate the statistical values.

Because it's not a single question of the regional statistics, the use of recalculation procedures after a territorial change has to be forced in all productions steps.

Let's have a look to the

F8: Role of a central Regional Classification Server

<C> We see a statistical production process A – e.g. population statistics – and produce a statistical report A.
<C> In another domain (e.g. hospital statistics) we find the equivalent process, producing a report B.
<C> For analytical reasons information of both reports has to be combined e.g. to calculate an index like hospital beds per 1000 capita.
<C> To be sure that in both domains the same regional classification is used, one has to provide a central regional classification server (RCS).
<C> The use of the RCS should be mandatory; assigning values to geographical units is called geocoding and has to be performed in all production processes.
<C> If it's recognized that the classification has changed, a recalculation of older values has to be performed.

The production units may support the maintenance of the RCS by reporting changes which are occur in collected data.
<C> This feedback enhances the quality of the RCS.

The RCS has to be integrated into the IT infrastructure of the NSO, as you can see in the next diagram

F9: You can see the components RCS, productions units and the regional statistics DB.
<C> The metadata base is shown, because the RCS is very close to Metadata base, some see the RCS as part of the metadata, like other classifications. (NACE, …)

<C> RCS info has to be fed into the production process, so that the uploading process into the Regional Database runs without problems.

<C> Nevertheless a quality proved, semi automatic uploading process checks the data against the RCS.

<C> The obligation to use the actual regional classification exists not only for internal production units but also for external data producer.

<C> Therefore the external data producer must have access to the RCS also.

How is this model implemented in a real world shows the next slide:

F10: The Solution in the Turkish NSO TURKSTAT

In TURKSTAT the RCS is for organisational reasons not seen as part of the Metadata Server. We see the components as described in the previous foils.

<C> The online products of the Regional Statistic are Table which can be composed by the user individually by theme, time and regional level. After displaying the table on screen, the table can be downloaded in EXCEL, PDF or HTML. It is also possible to see the regional distribution of selected variables on different levels on maps.

<C> Because a lot of basis statistic processes are performed centrally, where you need address information, the RCS is based on an nationwide Address Server.
<C> The original Address Database is maintained in connection with the population register in the Directorate General for Population and Citizenship, a subordinated administration of the ministry for interior affairs.
<C> The suggested information flow is the following:
Extraction of all changes of the National Address DB on a daily base to integrate into the TURKSTAT Address DB. This information gives clues which regions have changed and have to be adjusted.

<C> Also external data providers should use the National Address DB and the regional classification stored, to provide correct geocoded data.

So fare the description of the required infrastructure and the integration into an NSO environment.
Next it’s necessary to provide some information about functionality which is required to support the recalculation.

F11: Recalculation of values and indicators**.**

We have to distinguish between the values counted and the indicators derived from these values.

<C> Basic - Values
If surveyed / reported values are available, these values have to be aggregated from the original level to regional units according the new territorial division
<C> If surveyed / reported values are not available, the existing values have to be recalculated using weighting factors
These can be derived from other sources / information, which is preferred

If these additional sources not available, weighting factors derived from change of area or population can be used to approximate the change. These general weighting factors should be provided in the RCS. This will be discussed later.

<C> If new Basic Values have been recalculated, all Indicators depending on this values must be calculated from the basis values on all regional level

F12: This foil shows the recalculation process. We see on the left side the original values based on the regional classification of 1991 and in the middle part the recalculated / target values based on the regional classification of 1996. On the right hand the difference is shown.

To show what is required to do this recalculation, we will analyse the content of the RCS on conceptual level.

F13: Let's start with the description of the regions on one certain level.

The Standard Information which is mostly kept in an simple table, is a reference table of the region name and the region code used in the databases.

If histories are recorded for each region a lifetime is recorded, mostly in a start / begin / from date and an end / to date.

<C> If it's planned to provide weighting factors area or population, the assigned values have to be kept also.

F14: On the green background, the description of the provinces of TURKEY can bee seen.

The recording started on October 21 in 1990, the date of the 1990 census. Therefore all provinces exist at this date have a "From" date of 21-10-1990.
Let's have a look to the province ADANA in the first four rows. The region code is "01" and the name is "Adana".
The first change of this province happened on Feb 10th 1993, the second on April 9th 1995, and the third on October 25th in 1996. All dates are recorded in the column "From" date. One day before the lifetime of the old extension of the region expired, which is recorded in column "To" date. The last entry represents the actual region, which has an open end, marked as 31.12.2999.

In the regions "04 Ağrı" and "0 Antalya" the extension never changed. This is shown by the earliest From date (21-10-1990) and the open end date (31.12.2999)

In the column “Population” the equivalent value of the census 1990 is recorded, according to the territorial division. It can bee seen that the first two changes are very moderate, only some villages have been assigned to another province. The third change is significant (nearly 400.000 inhabitants got lost), the province was splitted up and a new province was created.

The columns on the right hand side are explained later.

Next, it’s important to record the required change information.

F15: This is done by five additional columns.

The date of change, the regional unit from where the changes go (From Unit) and where the changes went (To-Unit). The change is described in the columns “area” and “population”.

F16: These are clippings of the full table, showing the changes of Adana and where they are gone.

The first change was the reassignment of a small village called “Incirgediği” to the province “33 Icel”, affecting 258 inhabitants.

As we can see in the middle, the province Icel also changed, but with positive effects, resulting from the adopting “Incirgediği”. The population of Icel rose by 258 persons.

The second change affecting Adana, was the village “kaleboynu”, which affected 1312 inhabitants. This village was reassigned to the province “46 K Maras”, as can be seen in the table.

The third change affecting Adana, was the establishment of the province “80 Osmaniye”. Due to this Adana lost 384.104. The creation can be recognized. The province itself has no start population, because the end population is equivalent to the amount of changed persons.

Now we have seen the kernel structure of a RCS. The columns with the brown background only for a better understanding of the change, but need not implemented in a system.

Next is, to describe some of the products, which can be derived from the RCS.

F17:
The Status Table, the correspondence table and – as an exmple what the EU is claiming – The CONC MODA tables.

F18: The Status Table

This table contains a list of all valid regional units for a certain date. If I'm a data producer, I use this table - which is created on the fly upon request - to align to the classification in my data.

As shown this table shows all regional units valid on October 10th 1998.

Similar tables are requested from the member states by the EU. As shown in the next foils.

F19: This is a CONC table.

The CONC / MODA tables describe the development of the NUTS-Units, the EU wide standardized territorial division of the nations and the relation to national units. These are deliverables of the EU member states to EUROSTAT.

The MODA table is comparable with the "green background" information of the RCS, the CONC the change itself.

Codification of concordance between the old and new units of territorial structure (CONC)

The clip shows the changes of Adana.

The country code of Turkey is "TR", the table type "CONC". The column "nat-code before" and "nat-code after" contain the equivalent national unique identifier. "Year, Month and day" describe the day of change. The column NUTS3 contains the Turkish NUTS-Code on level 3 (Provinces) and the column "label" the native name of the province.

F20: This is a MODA table, which can be seen on the second column. The content I have shown, except of the Change Code.

The change code defines the type of change as:
A = Add
D = Delete
M = Modification of label or national code

For me this is a little poor information to work with. I miss the at least the area.
A more informative table has to be created from the RCS as a base of the recalculation,

F21: The correspondence table.

This is also a table created on the fly on request of the user. All changes in a period (e.g. 1-1-91 and 1.1.96) are listed. For practical reasons, all regional units which are not affected by a change are included too.

The descriptions of a change – like Adana – start with a row containing the values at the start date.
In the first row we see the source ID and – name and the number of inhabitants on the beginning. This row is indicated by "SB" and a Factor Pop is 100%.

The second row describes the first change, indicated by "M", 258 inhabitants are reassigned to the province Icel. The Factor-Pop is 0,0133% which represents the 258 inhabitants of the 1.934.907 total population.

The third row repeats the change information for the second change to K-Maras.

The fourth row shows the province "Adiyaman" without any change in this period.

The columns Source-ID, Target-ID and Pop-Factor are used for recalculation. The process of recalculation can be described as follows:

F22: Recalculation of values

This foil shows the steps of the recalculation. Left hand side is the source table. This table is joined with the correspondence table described before. The value of the source unit is multiplied by the Factor-Pop. An additional record is inserted for the target unit where the subtracted value is added to preserve the total.

The table is sorted by regional units as shown in the Intermediate table (see middle part).

The values of the Intermediate table are grouped on source id and summarized. The result can be seen in the target value.

F23: Summary

Changing regional classification has a significant influence on the values of time series.

To provide correct data comparable over time, retrograde data have to be adjusted to the actual regional classification.

Some provisions on data and methods are shown. The RCS which has to be provided centrally and all producers of statistical data are forced to use this classification.

F24: Acknowledgements:

TURKSTAT

USST

F25: Questions ?

<C> Thank you for your attention