Integrated Database of International Migration Statistics with a Particular Attention to Linking Data Sources

Éva Gárdos
Chief Professional Advisor
HCSO
E-mail: eva.gardos@ksh.hu

Annamária Sárosi
Professional Advisor
HCSO
E-mail: annamaria.sarosi@ksh.hu

Áron Kincses
Councillor
HCSO
E-mail: aron.kincses@ksh.hu

Eleonóra Nagy-Forgács
Chief Senior Councillor
HCSO

International migration statistics use several administrative data sources, among those the registers of foreigners handled by the Office of Immigration and Nationality, and the register of personal data and addresses of the Central Office for Administrative and Electronic Public Services. Since 2002, these data sources have been complemented by an extensive survey on people who acquired Hungarian citizenship.

In the current study the procedure of linking the databases into a single integrated system will be reviewed.

KEYWORDS:
International migration.
Data processing.

The users' needs for quick and quality statistical data on different aspects of the economy and society impose a permanent demand for new efficacious methods of data collection on the statistical authorities. Utilization of administrative records may result in quicker statistics with low costs and high quality.

The present paper deals with the combined use of data of registers and surveys for international migration statistics at the Hungarian Central Statistical Office (HCSO).

Following a long period of isolation, migration has become a process affecting the Hungarian society in the last third of the 1980s. The situation changed fundamentally at the end of the decade. The historic events created a situation in which Hungary became a transit country to the West, as well as a final destination for immigrants (*Juhász* [2003]).

The need for an established information system on international migration came up at different levels of decision-making, public administration and civil society at the beginning of the 1990s. The HCSO started to compile the international migration statistics in 1992, using the administrative databases containing related information. The regular publication of migration statistics began in 1993.

The first major applied data sources were the population register and the central alien register. These administrative datasets were among the very first ones available for the statistics at the individual level. Later on these data sources were augmented with data on foreigners having work permit and on refugees. Recently the data of the national health insurance and those of the tax authority were included as well in order to obtain more complex information on the foreign population living or working in Hungary.

As there was an increasing need for having more information on new Hungarian citizens, a complementary statistical data collection was introduced in 2002. It contains information on demographic characteristics, previous citizenship, economic activity, occupation and the main reason for applying for the Hungarian citizenship.

Beyond the domestic needs, the statistics have to take the requirements of international organisations (UN, OECD) and the EU into consideration. Following a long period of gentlemen's agreements on the community migration statistics, the Regulation (EC) No 862/2007 of the European Parliament and of the Council has been issued recently saying "Harmonised and comparable Community statistics on migration and asylum are essential for the development and monitoring of Community legislation and policies relating to immigration and asylum, and to the free movement of persons".

In Hungary, the Act on Statistics (XLVI of 1993) guarantees that the HCSO has a right to use and link the official administrative data sources for statistical purposes. Furthermore, the acts establishing the migration-related administrative registers allow the HCSO the utilization of the data as well, mostly at an individual level.

1. Statistical concepts and areas

The basic definitions used in the Hungarian international migration statistics are in accordance with those of the United Nations Recommendations on Statistics of International Migration that are applied by the Regulation (EC) No 862/2007 of the European Parliament and of the Council too.

The most important concepts used in the present description are as follows:

Immigration and immigrant. According to the previously mentioned regulation, immigration in the present sense means the action by which a person establishes his or her usual residence in the territory of Hungary for a period that is, or is expected to be, of at least twelve months, having previously been usually resident in another country. Immigrant means a person undertaking immigration. In the Hungarian practice a foreigner having a residence permit for a longer stay or an immigration permit, who spent at least one year in Hungary after entering or got a residence permit for more than one year, is considered an immigrant.

Residence permit. At the request of a foreign citizen staying in Hungary with a valid residence visa, the local directorate of the immigration office issues a residence permit in order to lengthen the period of staying.

On 1st January 2002, the differentiation between the long-term (at least 1 year) and temporary (less than 1 year) residence permits was cancelled.

Permanent residence permits

1. Immigration permit (before 1st January 2002): A foreign citizen might have been given an immigration permit if he or she had been staying in Hungary for at least 3 years continuously and their accommodation and livelihood were secured and there was no excluding reason according to the laws.

2. Settlement permit (after 1st January 2002) /*National or EC permanent residence permit* (since 1st July 2007):[1] According to the revisions of the Alien Act, immigration permit was replaced by settlement

[1] Act I & II of 2007.

permit on 1st January 2002 and by national permanent residence permit on 1st July 2007. A permanent residence permit may be granted to a foreign citizen if he or she has lawfully resided in the territory of the Republic of Hungary continuously for at least the preceding three years before the application is submitted, except the case when the reason of stay is studying.

According to the national needs and international recommendations the Hungarian migration statistics of a given year cover the following main areas:

a) flows of immigrants and emigrants,
  – foreigners
  – Hungarians
b) stock of foreign migrants in Hungary,
c) naturalized population,
d) work permits,
e) refugees and asylum seekers.

2. Data sources

Migration statistics are primarily based on two administrative data sources: the population register and the central alien register (databases of the Office of Immigration and Nationality). These datasets are linked for providing knowledge on the stocks and flows of migrants, both foreigners and Hungarians, and the naturalized population. Separately, two other data sources supply statistical information on refugees and asylum seekers, as well as foreigners having work permit in Hungary.

In the following description, the population register, the central alien database and a statistical full-coverage survey of new citizens will be discussed in detail, whilst these data sources can be considered as an integrated solid data set.

2.1. Central Alien Register

Foreigners are registered in the Central Alien Register (CAR). The Office of Immigration and Nationality is responsible for this database (*THESIM report* [2005]). All foreigners applying for or holding a visa, a residence permit, an immigration permit or a residence permit and their accompanying children are included.

On the basis of the CAR, statistics are produced for

– foreign migrants staying in Hungary, 1st January,
– foreign migrants entering or leaving the country in the given calendar year,
– those having a valid permit for staying in Hungary as immigrants in the given calendar year.

On 1st January 2002, a new act was issued that amended the rules of immigration, and on 15th February 2002 a new register was opened. The cases are maintained in the old register, but every new case is included in the new one. On 1st January 2007, a separate register was established for the citizens of EEA (European Economic Area) countries.

Data entries: name, mother's name, date, place and country of birth, sex, citizenship, family status, educational attainment, occupation, addresses and date of entry.

2.2. Register of personal data and addresses (Population Register)

The major purpose of establishing the population register was providing reliable information on the names and addresses of the population for elections and referendums. It contains the most important data for identifying persons and separating different groups.

The Hungarian population register includes information on the following categories:

– Hungarian citizens having permanent residence (domicile) in Hungary,
– Hungarian citizens having permanent residence abroad (living abroad or living temporarily in Hungary) upon their request,
– foreigners with permits for permanent residence (including recognised refugees and stateless people),
– EEA citizens with residence permit.

Foreigners staying in Hungary temporarily (for example foreigners with a residence visa or a "temporary" residence permit, foreigners with a certificate entitling to temporary stay, foreign diplomats and asylum seekers) and the overwhelming majority of the Hungarian citizens staying temporary abroad are not included because they abstain from reporting that they leave the country. Moreover, the population register covers very limited information on individuals, not even allowing us to ex-

plore the basic characteristics of the migrant population. This is the reason why the Hungarian migration statistics are based on several administrative and statistical data sources and the population register is only one of them.

The information included in the population register is the following: name, mother's name, date and place of birth, country of birth, sex, citizenship, address(es), date of registration, date of log out of the register, cause of registration, cause of log-out, family status, and date of acquisition of the Hungarian citizenship.

2.3. Statistical survey on people who acquired Hungarian citizenship

After getting the citizenship, all new Hungarian citizens are asked to fill in statistical forms on themselves and on their minor children who acquired the Hungarian citizenship together with them. Filling in the questionnaire is voluntary. The respondents are asked to provide information on personal data (name, mother's name, date of birth, place of birth, country of birth, address), on demographic characteristics such as sex, family status, number of children, and on mother tongue, previous citizenship, educational attainment, reason of application, economic activity and occupation before entering Hungary and currently, on the date of the acquisition of the Hungarian citizenship. The questionnaires are distributed to and upon completion collected from the local governments by the HCSO.

3. Control and data editing

The controlling and data editing processes differ in the case of the Central Alien Register, the Population Register and the statistical survey on people who acquired Hungarian citizenship.

3.1. Central Alien Register

The data editing process in the case of the Central Alien Register has six main steps:

1. The Central Alien Register is a complex dataset containing records of each measure made in reference to a person. Combining data of records belonging to the same person, a new record of the necessary data must be produced. In order to merge the different records of a given person, a key variable (based on the main identifying variables such as the name and the date of birth) shall be used.

2. The register includes three sub-registers in accordance with (the changes of) the legal rules:

– Register of foreigners registered by the immigration authority before 15th February 2002,

– Register of EEA-citizens,

– Register of third-country nationals.

However, these sub-registers use different coding rules for certain common variables. Since the statistical use necessitates a single code list for a variable, the original different code lists are converted into the one used by the statistics.

3. Foreigners with permanent residence permits are included in the Population Register as well. In the case of the common variables, the missing or out-of-date information in the Central Alien Register can be substituted by that of the Population Register. To achieve this, the two datasets are merged using a record linkage technique.

4. Whenever it is possible, standard classifications (such as International Standard Classification of Occupations, ISCO) are used instead of the register's own classifications.

5. Automatic corrections make the different dates relating to one person coherent. Furthermore, information on educational attainment, occupation, family status and age is also harmonized.

6. According to the data describing the present status (type of permit, duration of stay in the country, date of entry) of the foreign person, statistically relevant groups of migrants are formed.

3.2. Population Register

Only the completeness of the records selected for the purposes of international migration statistics is checked. It is investigated whether the number of records in the various groups of the foreign population fits to that of the previous years, taking into account the information of other data sources.

3.3. Statistical survey on people who acquired Hungarian citizenship

Control and editing of survey data consist of four steps:

1. Checking if all completed questionnaires have been received by the HCSO. In 2002, when the statistical survey was introduced, almost

three fourth of the new Hungarian citizens filled in the statistical form (*Kincses* [2003]). It may be considered quite a good proportion taking into account that answering is voluntary. In the following years the response rate was even somewhat higher.

2. Check of the validity of code numbers and the logical consistency of answers. In the case of mistakes, correction is made using the available information. The response rate of the individual questions is rather high. It is the lowest as regards the educational attainment, where it almost reached 90 percent.

3. The survey data are compared with the data of the registers concerning the naturalised people in order to control whether the two datasets can be statistically considered the ones provided by the same population.

4. The survey data are integrated with the related records in the population register using a statistical technique.

3.4. Linking techniques

The datasets are linked using two approaches: a record level- and a statistical linkage.

Record level linkage of the Central Alien Register with the Population Register

The two administrative databases are integrated with a one-to-one record linkage technique. This data linkage serves two aims related to the common variables of the two registers for foreigners having permanent residence permit. One of the purposes is to update the variables in the immigration register and the other is – in case of deficiency, mistake or inconsistency – to correct them, using data of the population register, which is based on registrar's reports.

A key variable is constructed by merging the family name, the given name and the date of birth. In order to have an unambiguous personal identification code, the names are curtailed of accents. In the first step, the records of the two datasets having the same key variable are linked to each other in an automatic matching process. Finding the pairs in the rest of the cases is done manually using further information if necessary.

Statistical linkage of the Population Register with the statistical survey

The combination of the administrative data with the survey data is divided into two parts. First, it is checked whether the two data sources provide information on

the same population neglecting the acceptable errors of the statistical survey. Secondly, a statistical data linkage is carried out.

As a preparatory action the closeness and stability of the relationship between the two databases are measured with two statistics.

a) The linear correlation coefficient is as follows:

$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$, where x_i, and y_i are the values of the variables in the two data sources, while $\bar{x}$, and $\bar{y}$ are the related means and $-1 < r_{xy} < +1$.

b) The elasticity coefficient is as follows:

$E = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \frac{\bar{x}}{\bar{y}}$. It indicates how much change of the variable *y* is caused by 1 percent change of the variable *x* in the given relationship (*Hunyadi – Vita* [2004]).

Three common variables are used for the comparison: previous citizenship, county of the place of residence and age. These variables are rather different with respect to the number of categories and the coverage ratio of the individual groups.

1. In the case of previous citizenship only the European countries are considered due to the low number of cases outside Europe and to their lower reliability. The following five categories are used for citizenship: Romania, countries of former Yugoslavia, Ukraine, other EU member countries and, the rest of Europe.

On the basis of both coefficients the two datasets are proved to be highly consistent.[2]

2. Considering the frequencies by the county of the place of residence, the ratio of the register cases covered by survey records is very different by counties (20 including the capital). It might be explained by the increased mobility of the migrant population that they declare

[2] The coefficients for 2002 were as follows: $r_{xy} = 0.9971$ and $E = 1.11$.

the official address in the register, but in the survey they indicate the actual address or the one where they plan to move soon. The coefficients show somewhat less correspondence than those of citizenship.[3]

3. Looking at the differences of the two datasets by the distribution among the eight age groups (0–14, 15–19, 20–24, 25–29, 30–39, 40–49, 50–59, 60–X years), it can be stated that the coverage ratio is the lowest for the 0–14 year old children. The questionnaire is filled in for less than half of the children in the register. It is due to the fact that the parents are less willing to fill in the form for their children than for themselves. The closeness of the relationship in this case was the lowest.[4]

Based on the foregoing, the two datasets are not independent. We might suppose that the records of the two data sources relate to the same population. This checking procedure is repeated each year.

As the aim of the statistical integration is to produce two-dimensional frequency or contingency tables, there is no need for exact matching of the records of the two data sources. Furthermore, the statistical survey will never cover the 100 percent of the cases included in the register that hinders the implementation of a simple one-to-one record level integration. It seems reasonable to use the RAS method for linking the two data sets rather than apply a record linkage (*Stoyan–Takó* [1993]; *Kincses* [2003], [2004]).

Let us consider a two-dimensional table of the statistical survey. One of the variables is common in the two data sources. It can be supposed that the variable of the column is the common one, while that of the row is included only in the data set of the statistical survey. The elements of the table are as follows:

$$\begin{array}{cccccc|c} a_{11} & a_{12} & . & . & a_{1n} & a_{1.} \\ a_{21} & . & . & . & a_{2n} & a_{2.} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ a_{m1} & a_{m2} & . & . & a_{mn} & a_{m.} \\ \hline a_{.1} & a_{.2} & a_{.3} & . & a_{.n} & a \end{array}$$

[3] Although the proportions varied between 42 percent and 100 percent in 2002, the coefficients showed close relationship, r_{xy} = 0.9851 and E = 0.9278.

[4] In 2002, the average age was 37.93 years in the register and 40.58 in the statistical survey, r_{xy} = 0.9503 and E = 0.8927.

a_{ij} denotes the element in the cross of the row *i* and the column *j*.

$a_{.j} = \sum_{k=1}^{m} a_{kj}$, for any $j \in \{1,2,.....,n\}$, and $a_{i.} = \sum_{l=1}^{n} a_{il}$ for any $i \in \{1,2,.....,m\}$,

and $a = \sum_{b=1}^{m} \sum_{c=1}^{n} a_{bc}$.

The RAS method modifies the elements of the previously mentioned table in a way that the inner distributions will remain the same and the table will fit to the register data too.

As the first step, the column values will be changed to the ones in the register. The new values will be denoted as $b_{.j}$ (*j* = 1,2,….,*n*) and the grand total will be *b*.

.....	.	.	.	.	.
.....	.	.	.	.	.
.....	.	.	.	.	.
.....	.	.	.	.	.
.....	.	.	.	.	.
.....	.	.	.	.	.
$b_{.1}$	$b_{.2}$	$b_{.3}$	.	$b_{.n}$	b

Let us change a_{ij} to a'_{ij} so that the column sums will remain and the change of the elements will be proportionate. Thus, the new elements will be as follows:

$a'_{ij} = \frac{b_{.j}}{a_{.j}} \cdot a_{ij}$ for any $i \in \{1,2,.....,m\}$, and any $j \in \{1,2,.....,n\}$, and they fulfil the following equation: $\sum_{i=1}^{m} a'_{ij} = \sum_{i=1}^{m} \frac{b_{.j}}{a_{.j}} \cdot a_{ij} = \frac{b_{.j}}{a_{.j}} \sum_{i=1}^{m} a_{ij} = \frac{b_{.j}}{a_{.j}} \cdot a_{.j} = b_{.j}$. This means that the column sums are as expected and the inner elements are proportionate.

The row sums will be as follows:

$b_{i.} = \sum_{j=1}^{n} a'_{ij} = \sum_{j=1}^{n} \frac{b_{.j}}{a_{.j}} \cdot a_{ij}$ for any $i \in \{1,2,.....,m\}$, and any $j \in \{1,2,.....,n\}$,

The sum of the totals equals to the value of the grand total:

$$\sum_{i=1}^{m} b_{i.} = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{b_{.j}}{a_{.j}} \cdot a_{ij} = \sum_{j=1}^{n} \frac{b_{.j}}{a_{.j}} \left(\sum_{i=1}^{m} a_{ij}\right) = \sum_{j=1}^{n} \frac{b_{.j}}{a_{.j}} \cdot a_{.j} = \sum_{j=1}^{n} b_{.j} = b\,.$$

In the way described previously a projection was carried out that keeps the inner relationships invariant combining the pieces of information of administrative and statistical data sources.

4. Summary and conclusions

In the recent period not only the statistical use of the administrative data sources but also their integration has been coming to the front. It is caused by the increasing demand for quick and cost-effective data collection methods that decrease the burden of data providers, and by rising awareness of the potential benefits to be gained through matching disparate databases.

Migration statistics is one of the domains, which must rely on administrative data sources. Migration policy should be based on an information system comprising data on immigrants and their integration in many aspects. Such a dataset shall be the result of merging a number of separate registers and statistical surveys. In order to link them, a unified set of identification data should be used. The demand is very similar to the one of a register based census that has become more frequent in the world.

The better accessibility of administrative data, the more frequent use of them and the closer co-operation among the stakeholders will contribute to a unified system of migration data of high quality.

References

OFFICE JOURNAL OF THE EUROPEAN UNION [2007]: *Regulation (EC) No 862/2007 of the European Parliament and of the Council of 11 July 2007 on Community Statistics on Migration and International Protection*. Vol. 50. L 199. pp. 23–29. http://eur-lex.europa.eu.

HUNYADI, L. – VITA, L. [2004]: *Statisztika közgazdászoknak*. KSH. Budapest.

JUHÁSZ, J. [2003]: Data System Description-Hungary. Compstat. www.compstat.org

KINCSES, Á. [2003]: *A magyar állampolgárság megszerzésének statisztikájához tartozó adatforrások leírása.* Working Paper.

KINCSES, Á. [2004]: *A magyar állampolgárság megszerzése, 2002–2003*. KSH. Budapest.
STOYAN, G. – TAKÓ, G. [1993]: *Numerikus módszerek 1*. Typotex Kiadó. Budapest.
THESIM REPORT, HUNGARY [2005]: www.cefmr.pan.pl/docs/thesim_report_hu.pdf
www.ksh.hu