Administrative data and survey data on electricity use in Hungary

Akos Gerencser[1]

[1] *Hungarian Energy and Public Utility Regulatory Authority, Budapest, Hungary; gerencsera@mekh.hu*

Abstract
The official energy statistics is compiled solely by Hungarian Energy and Public Utility Regulatory Authority as Other National Authority (ONA) in Hungary.
The annual energy statistics covers energy use statistics broken down by energy products (including electricity) and activity sector by NACE groups. Data are traditionally compiled based on an end-user sample survey.
On the other hand, electricity suppliers have also recently started providing detailed data on electricity supplied to end-users, which can be used for statistical purposes by matching via the General Business Register.
The paper briefly describes the method of the administrative data collection, the overall quality of data obtained compared to survey data, the potential benefits and limits of using these administrative data for statistical purposes.

Keywords: administrative data, energy use, Other National Authority

1. Introduction

The Hungarian Energy and Public Utility Regulatory Authority (HEA) is the regulatory body of the energy and public utility market in Hungary and also part of the European Statistical System as Other National Authority (ONA) since 2012. Thus, HEA compiles the monthly and annual energy statistics including the energy balances according to Regulation 1099/2008/EC of the European Parliament and Council. As part of the annual energy balances, detailed data are required on final energy consumption broken down by both energy products and end-use sectors (pre-defined NACE groups). This paper is focusing on the statistical data collection and processing methods applied for producing the annual final consumption data and the new approach of using administrative data for electricity consumption via data linking techniques.

2. Statistical survey

There are no universal data collection strategies defined in the international recommendations for energy statistics (IRES, UN, 2016, see para. 7.15.). The implementation guidelines for IRES, the draft Energy Statistics Compilers Manual (ESCM) includes further details for data collection in Chapter 4b and has identified four potential data sources each having its advantages and disadvantages namely statistical survey, administrative data, modeling and metering.

In Hungary there is an annual statistical sample survey system implemented, which is surveying energy end-user organizations directly by targeted questionnaires for all types of energy consumption including electricity as well. The data providers are selected based on the frame of the General Business Register (GBR) maintained by the Hungarian Central Statistical Office (HCSO) and obtained by HEA once a year. There is a cut-off threshold defined by the number of employees above which all units are selected irrespective of their NACE activity and there is a stratified random sampling applied below the threshold. Micro enterprises are not surveyed. The sample size is heterogeneous depending on the sector of activity. The inclusion probability is higher in energy intensive activities (e.g. Industry). Altogether, there are ca. 3800 entities surveyed from Industry; 700 from Agriculture, forestry and fishing; 1000 from Transport and 2000 organizations from Commercial and public services sectors obliged for data reporting. In recent years there have been significant IT developments introduced in the data collection system, both to reduce the burden for data providers, and also to simplify and improve data processing. Nevertheless, the processing of data including validation still requires significant resources, as reporting errors are quite common (e.g. misreporting values due to wrong level of unit measurement, renting transactions, etc.)

3. Methodology of exploiting administrative data

HEA is entitled to collect data from energy branch as supervising authority. As most electricity is delivered through the electricity distribution system to end-users, electricity distribution system operators (DSO-s) are potential data providers for electricity final consumption. This administrative data source has traditionally been used for measuring

aggregated annual residential consumption. As DSOs have a rich administrative data source about electricity delivered to customers both from geographical point of view through point of delivery (POD) information and can identify customers at individual level, it was found logical to extend using their administrative data for energy statistics purposes.

As it was described in the previous chapter, HEA has access to the General Business Register of HCSO. The ID used in the General Business Register in Hungary can be directly derived from the tax-ID, which is also used in general as VAT-ID (value added tax ID). The key idea of making use of the administrative data of the DSO-s was to ask them to identify their customers not only by Customer ID but also by VAT ID, which they should have available in their invoicing systems and consumption data can be then linked to the dataset of GBR.

4. The linking exercise and results

The six electricity DSO-s of Hungary were requested in 2015 for the first time to provide two datasets each on 2014 annual distributed amount of electricity to end-users.

The first dataset contained delivered amounts in megawatt-hour (MWh) to households broken down by settlement and number of customers in the specific area (the geographical extension of the existing report on residential consumption).

The second dataset contained delivered amounts in MWh to non-household customers by settlement, VAT-ID and NACE. (Reported NACE could be used as a fallback option, if the VAT-ID could not be matched with GBR.)

DSO-s were assisted in the first phase before sending the real data, by checking the validity of the partner VAT-ID list generated from their system. The partner lists yet contained also some ID-s which could be matched with the GBR, but the entities have already stopped their operation according to GBR. Nevertheless, it was presumed in these cases that the NACE activity of the electricity consumption associated to the non-operating unit in the GBR has not changed.

Also, some minor adjustments had to be made in the geographical dimension by cross-checking postal ID-s and settlement names and also the service providing area of the given DSO.

After all, the matching exercise can be considered as very successful, i.e. the matching rate of the reported and adjusted data with the GBR is very high, as illustrated in the below table:

	No. of reported entities	% of matched entities	Total reported non-household MWh	% of matched MWh
DSO 1	19 907	98,54%	2 909 620	99,94%
DSO 2	18 163	87,57%	2 746 475	96,72%
DSO 3	29 220	98,66%	5 463 128	99,77%
DSO 4	40 055	99,08%	6 340 123	97,12%
DSO 5	11 274	99,00%	4 181 231	97,94%
DSO 6	18 065	98,70%	2 576 421	99,89%
TOTAL	**136 684**	**97,32%**	**24 216 998**	**98,45%**

5. The comparison of survey and administrative data

The GBR matched administrative data of DSO-s provide an alternative top-down data source compared to the bottom-up statistical survey data for the NACE group breakdown of electricity consumption.

In the below table 2014 annual non-household electricity use data are compared by NACE groups as estimated by the statistical survey and derived by matching DSO-s data with GBR. (Please note, that data are activity based both in admin and stat. survey data and not adjusted here according to energy statistics methodology, e.g. Rail includes all consumption by entities with NACE 491 or 492, irrespective of electricity used for transport or in buildings):

Sector	NACE group	Admin data (MWh)	Stat. Survey (MWh)	Bias
Industry	Iron and steel	656 152	591 796	10,87%
	Chemical and petrochemical	2 503 208	3 197 173	-21,71%
	Non-ferrous metals	406 207	389 358	4,33%
	Non-metallic minerals	1 043 787	1 203 512	-13,27%
	Transport equipment	1 062 167	1 465 869	-27,54%
	Machinery	1 866 674	2 458 510	-24,07%
	Mining and quarrying	81 479	86 508	-5,81%
	Food, beverages and tobacco	1 801 362	2 082 295	-13,49%
	Paper, pulp and printing	692 442	798 053	-13,23%
	Wood and wood products	221 548	244 932	-9,55%
	Construction	202 155	260 534	-22,41%
	Textiles and leather	158 213	199 527	-20,71%
	Not elsewhere specified	1 332 179	1 697 584	-21,52%
Transport	Rail	977 829	899 537	8,70%
	Pipeline transport	10 260	12 097	-15,19%
	Road	311 866	410 430	-24,01%
	Domestic navigation	1 639	1 950	-15,92%
	International aviation	43	745	-94,21%
Other	Commercial and public services	8 318 663	7 466 788	11,41%
	Agriculture/forestry	784 716	796 538	-1,48%
	Fishing	11 958	15 417	-22,44%

In general, the results of the two data sources are largely comparable and thus the linking exercise can be considered as very successful. In the most NACE groups the bias is around or less than 20%. The consumption is usually higher according to the statistical survey than in the admin data in the Industry and Transport sectors, while vice-versa for Commercial and public services. The main reason for that is supposed to be the re-sale of electricity by real estate transactions or in industrial parks, where Company 'A' is the single customer of the DSO registered in Commercial and Public Services (e.g. NACE 'L' - Real estate activities) by GBR and is different from the actual end-user Company 'B' operating in Industry sector and renting the facilities from Company 'A'. Yet, according to the GBR linked admin data 1,5 million MWh consumption is allocated to NACE section 'L' belonging to Commercial and Public Services, most of which is actually consumed probably in other NACE groups and/or sectors.

By analyzing the data in detail, some other reasons for bias were also revealed, such as the self-generation and subsequent on-site use of electricity in certain industries (e.g. Chemical and petrochemical) which is reported in survey but missing from admin data. Also some large consumers are not taking electricity from DSO-s, but directly from the electricity transmission system, which is also missing from admin data for the time being.

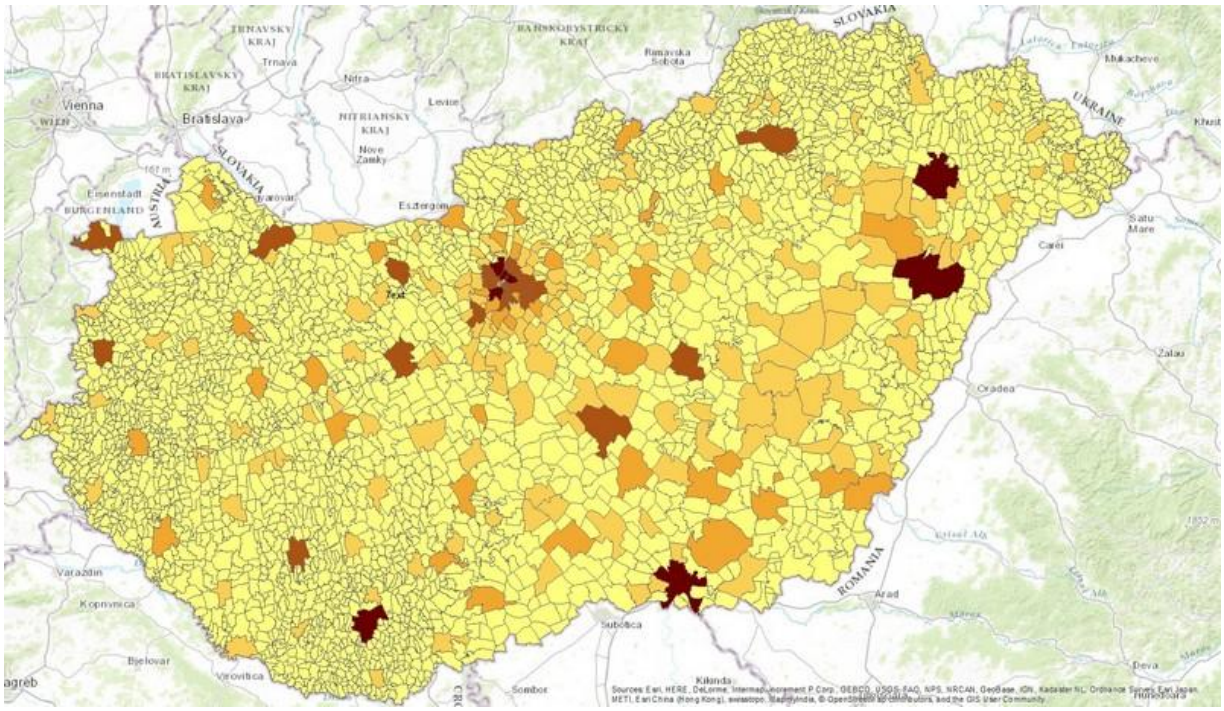
In that sense, statistical survey data are currently considered more reliable for producing NACE breakdown of electricity end-use and cannot be directly replaced by admin data for many, even large consumers, as bottom-up approach is showing a different picture compared to top-down admin data.

6. Conclusion, way forward

The successful linking to GBR enabled producing not only detailed NACE sectoral data, but also the analysis of electricity consumption by size-class, income, value added, etc. Compared to the sample size of ca. 8000 organizations of the statistical survey, administrative data are available for (almost) the total population without additional costs. Detailed data can be used very effectively by the calculation of various energy efficiency indicators, comparing the electricity consumption across the country, in different activities and size classes. Comprehensive data for small and medium enterprises are extremely valuable, as they are usually not included in large numbers in the statistical surveys.

The raw quality of administrative data is expected to be better than statistical reports of end-users and thus administrative data can be used in the validation of the reported survey data.

One of the main user needs which initiated the linking exercise of administrative data was to provide reliable small-area breakdowns, compared to statistical survey which cannot be representative at such detailed level. The below map is a first result illustrating small area results, showing the density of annual residential electricity consumption by settlements:



As an overall conclusion of the linking exercise, admin data turned out to be very useful as a complementary data source and the investigations should be further continued. The linking exercise is also extended to natural gas with the same methodological principles, but some more efforts should be made to improve the coverage and quality of linking with GBR.

For the moment it is concluded, that admin data even linked perfectly to the GBR cannot completely substitute the costly statistical surveys for end-use energy consumption, as surveys provide an overall information from a single data source on the consumption of all energy products and not only for selected grid based products such as electricity or natural gas, and because of the bottom-up approach are more tailor-fitted to the methodological requirements of energy statistics.

7. References

Unites Nations (2016), International recommendations for energy statistics (IRES), White cover publication, N.Y.

Oslo Group (2014), Energy Statistics Compilers Manual, Chapter 4b - Data sources and data collection, Draft online version