

Possibilities of spatially detailed statistical analysis in grid level integrated datasets¹

Ákos Jakobi

(Hungarian Central Statistical Office)

akos.jakobi@ksh.hu

In the emerging data era, demand for increasingly detailed information is steadily growing, hence providing the option to mirror processes of the economy and society more adequately. Spatially detailed (or granular) datasets have continuous importance in statistics, since many social or economic phenomena vary significantly across space. Aggregated data can hide important local patterns, while granular spatial data allows analysts to understand and model these variations more accurately. The main potential of detailed data lies for example in the possibility of revealing local heterogeneity, since national, regional, or city-level averages often mask substantial differences within areas. Granular spatial data, therefore, helps statisticians identify these local differences and avoid misleading conclusions. Additionally, better policy design and resource allocation could be achieved with the detailed view, especially at evidence-based decision making. The spatially detailed thinking is also supported by the general advancement in geospatial technologies and the expansion of procedures that generate data at fine spatial scales. Moreover, there is an improved need for better models of spatial dependence, since – according to Tobler’s well known first law of Geography – nearby locations generally tend to be more similar than distant ones (Tobler 1970). In addition, there is an increasing demand for precision decision making, namely, to answer the question: “what is happening exactly where?” Businesses, governments, and researchers need geographically specific insights rather than broad averages, formulating an increasing task for official statistics.

On the contrary, there are also challenges that modern spatial statistics must tackle when dealing with granular spatial data. The main problems are generally originated from massive data volumes, privacy concerns or computational complexity, just to mention some.

Numerous new data sources are now available that extract highly detailed information from the “big cloud”. Although such disaggregated datasets can usually be generated independently, their true added value lies in their integration with other sources.

In response to exploit the advantages, the Hungarian Central Statistical Office (HCSO) has prioritized the multi-purpose use of data (the enhancing of interoperability), the benefits arising from structuring and linking datasets (creating a form of “data space”), as well as the interpretation and visualization of detailed data to ensure accessibility and usability. A new data ecosystem grounded in spatial granularity and interoperability is being developed to enhance both the relevance and the utility of official statistics. To address these challenges, the

¹ Paper presented at the 5th ISI Regional Statistics Conference, 3-5- June 2026, Valletta, Malta

Hungarian Central Statistical Office has initiated the Virtual Hungary pilot project, which seeks to organize the most detailed datasets into a fully integrated system.

The aim of the Virtual Hungary project is to create a synthetic nationwide socio-economic information system that provides continuous and almost real-time data for decision-makers on the actual and current state of the economy and society. The simulation that runs behind the synthetic system practically regenerates the state characteristics on a monthly or, in some cases, weekly basis, providing the chance to formulate a digital statistical twin on the country. Additionally, as a digital twin, the development will make it possible to run preliminary scenario models to support policy preparation studies and even legislative processes.

From a spatial perspective, as the data representation is carried out at the elementary level, the data will be queryable for any desired aggregation (e.g., territorial or social-group-based aggregates). Among the possibilities, the grid, as a spatial level proved to be one of the best for integration. Grid level integrated datasets are useful not only because they offered an opportunity to bridge the gap between detailed spatial information analysis and sensitive data disclosure, but also because of opening completely new avenues of spatial analytical possibilities. Examples like pseudo-raster operations, such as spatial moving average or convolution statistics, can better be performed on grids.

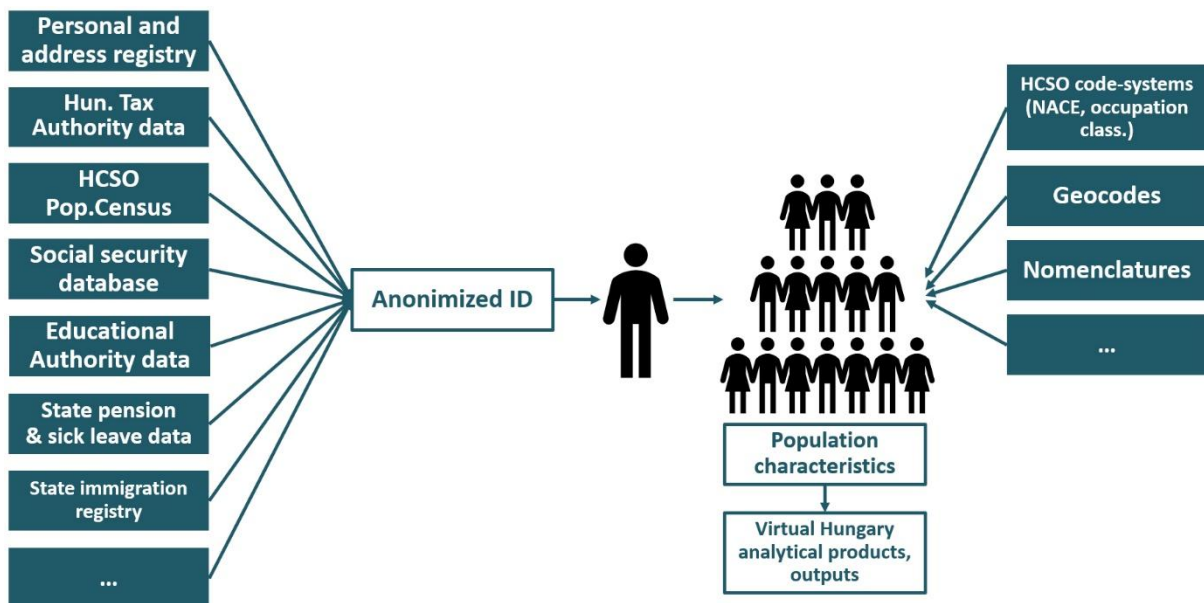


Figure 1. General framework of the Virtual Hungary statistical system (based on personal input datasets)

Beyond the analytical possibilities offered by examining the individual internal indicators of each separated dataset, the most interesting and novel analytical opportunities primarily arise from exploring the cross-domain relationships enabled by linking different data domains. The left-hand side of the figure (*Figure 1*) provides an illustrative example of selected Virtual Hungary data domains (subsets of the database), where the possibilities for data linkage are enabled by the anonymized Virtual Hungary IDs available at the individual level. The right-hand side of the figure shows selected sub-databases containing the classifications,

nomenclatures, and spatial identifier tables used by the HCSO. Although the complex data tables generated through cross-queries are at the individual level, their analysis can only be conducted at the level of population characteristics. In other words, Virtual Hungary analytical products and outputs are generated in forms that do not allow the identification of individuals.

Figure 2 shows, that individual-level data can be aggregated to a regular grid, enabling the spatial representation of key demographic and socio-economic characteristics. These include demographic characteristics as well as personal wage data, which can then be analysed jointly to explore the relationships between population subgroups and their average wage levels. By interlinking these different dimensions of individual-level information, it becomes possible to derive detailed local summaries that capture the socio-economic characteristics of specific areas. This approach thus provides a spatially consistent framework for analysing population composition and income patterns at a highly disaggregated level.

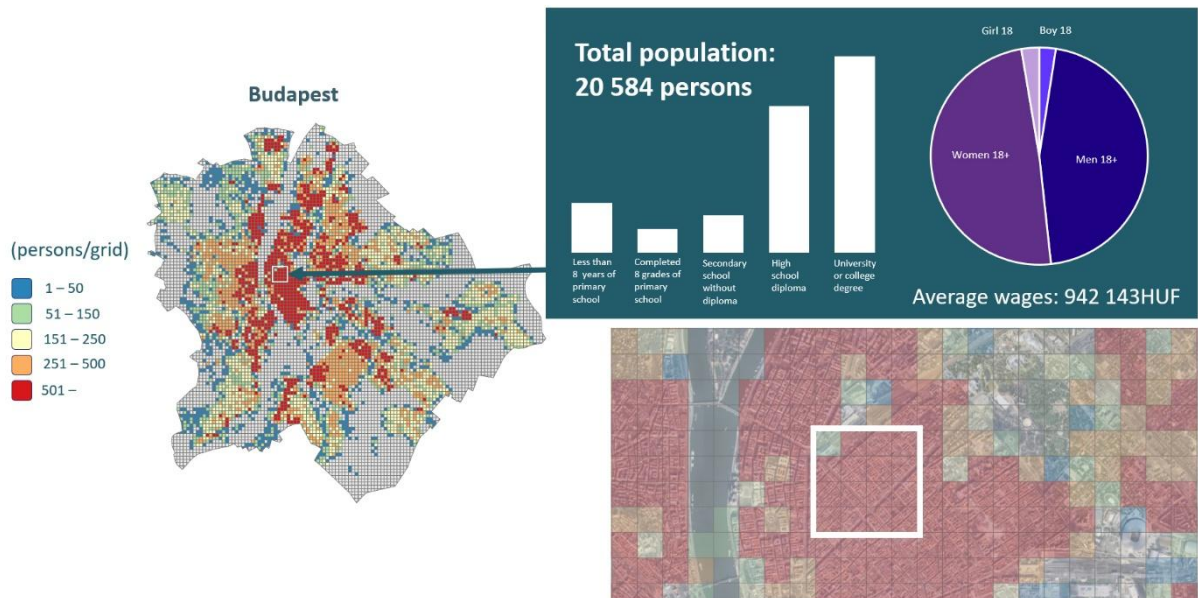


Figure 2. Example of a grid-level aggregated outcome for local summaries derived from interlinked data

Another major group of the Virtual Hungary data domains consists not of individual-level (personal) data, but corporate data, available at the level of enterprises. These data domains therefore cannot, in themselves, be linked to the individual-level databases described above (unless individual-level information can inherently be identified within a particular part of the data domain). The left-hand side of Figure 3 illustrates the datasets available at the enterprise (organizational) level, while the right-hand side again shows selected sub-databases containing the classifications, nomenclatures, and spatial identifier tables used by the HCSO, in this case pertaining to organizations. In such cases, the database linkages required for cross-queries can be established using organizational VIMA IDs. The resulting query and analytical outputs provide organizational population characteristics and thus, here as well, information at the group level.

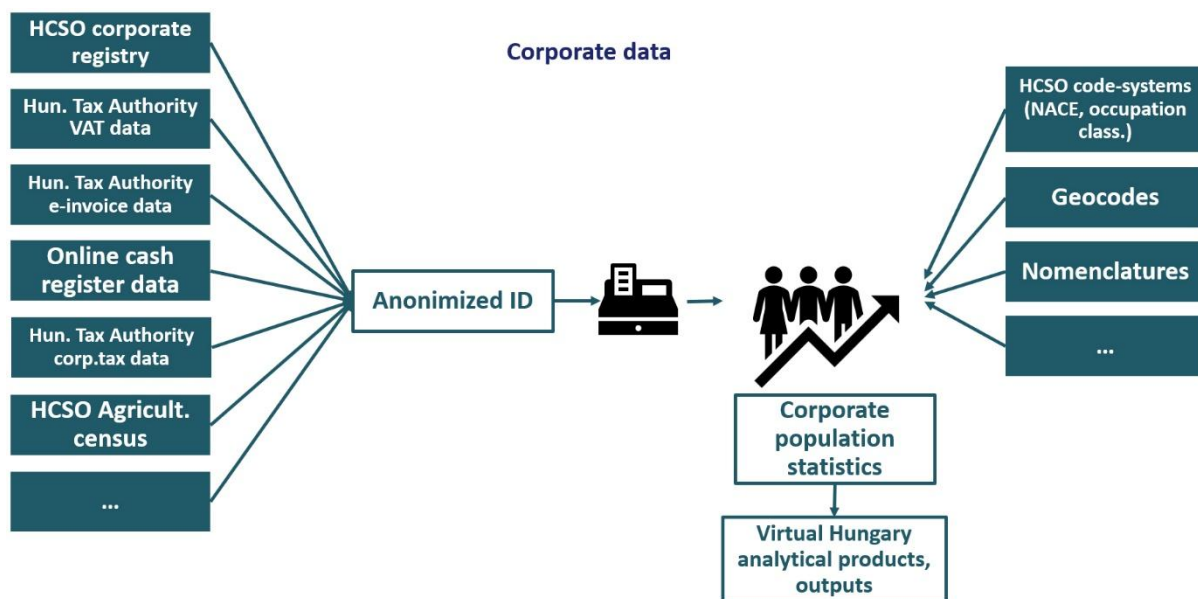


Figure 3. General framework of the Virtual Hungary statistical system (based on enterprise-level datasets)

Finally, as a special use case, individual-level and enterprise-level data domains can be used jointly for certain analyses. Figure 4 illustrates such a complex case, in which the example task was to determine which microregions' wage incomes would be most affected by a hypothetical 10% workforce reduction in the automotive industry.

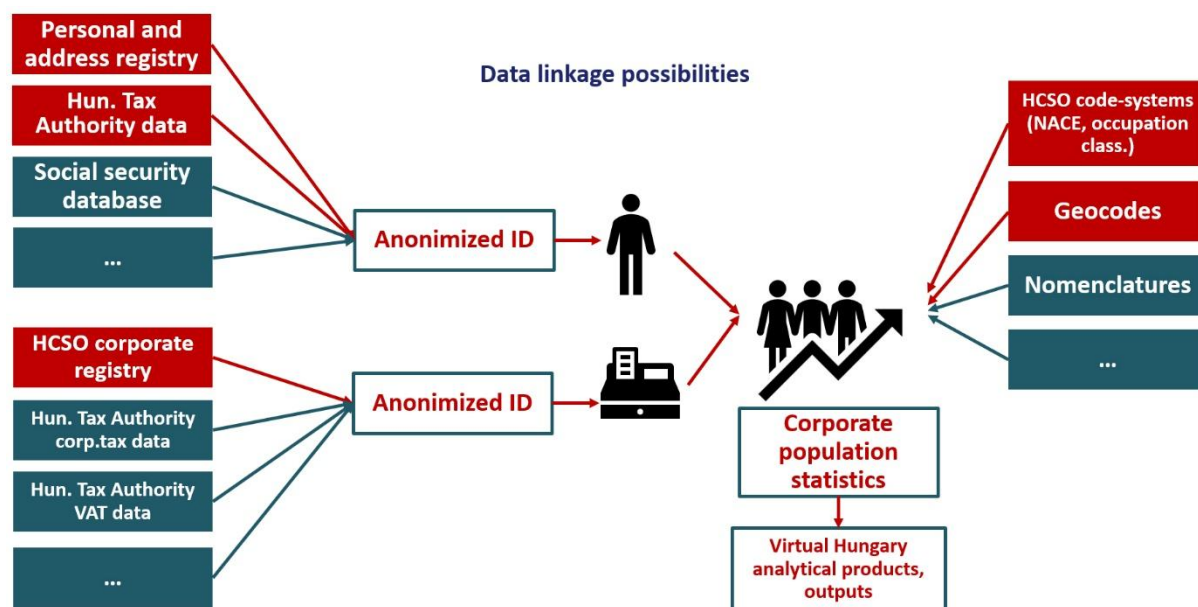


Figure 4. General framework of the Virtual Hungary statistical system with linkages between individual-level and enterprise-level data

To answer this question, it was necessary, on the one hand, to identify the location of the enterprises (at least at the microregional level) and determine their sectoral classification and, on the other hand, to establish the regional wage profile, which was based on individual-

level personal income data and residential address information. The integration of sources originating from different domains therefore made it possible to perform highly complex queries as well.

In conclusion, it can be stated that such interlinked, fine-grained information has so far remained largely inaccessible to the analytical community. The integration of detailed spatial data represents not only an opportunity but also an important task for official statistics. To fully exploit the potential of such data, a key path of implementation is the continuous expansion and development of data linkage capabilities, enabling the integration of increasingly diverse and detailed data sources. At the same time, official statistics are undergoing a continuous shift towards algorithm-based analytics and services, creating new possibilities for processing, interpreting, and disseminating complex datasets. These developments ultimately provide opportunities for the creation of new statistical analyses and products that can offer more detailed, timely, and spatially disaggregated insights into socio-economic processes.

References:

Tobler, W. A. (1970) Computer Model Simulating Urban Growth in the Ditroit Region. *Economic Geography*, 2, 234-240.