doi.org/10.26398/IJAS.0033-004

ADMINISTRATIVE DATABASE AND OFFICIAL STATISTICS: THE CASE OF THE REAL ESTATE ANALYSIS

Marini Caterina, Nicolardi Vittorio[1]

Department of Economics and Finance, University of Bari Aldo Moro, Bari, Italy

Abstract *The utilisation of administrative data in the official statistics is still far to be entirely resolved. Many research outcomes are mainly based on survey and sample data and very rarely integrate nonstatistical data. This paper faces the analysis of the real estate phenomenon that still suffers the dearth of an omni-comprehensive database of all information that derives from independent PA offices, also dealing with Big Data issues. To create the full information Harmonized Real Estate Database, a method to perfectly align all administrative and statistical data has been designed and employed. The empirical evaluation of the real estate phenomenon in the Italian city of Bari has demonstrated the enormous potentialities of the harmonised database for new socioeconomic analyses and challenging statistical advancements.*

Keywords: *Administrative Data, Big Data, Real Estate Economy, Harmonised Database*

1. INTRODUCTION

Since the beginning of the 21st century, when the utilisation of Internet connection was starting to change the communication systems and the way of dealing with the globalised information, the scientific debate over the new opportunities of the research frontiers internationally involved all fields. The important Information Technology explosion that simultaneously occurred and profoundly transformed all sectors of the economy in managing the corresponding activities alongside the voluntary generation of public data on the network represented the beginning of a new era that will extraordinarily benefit from this revolution. The global computerisation of the bulk of activities in the public administration, industries, and private enterprises has revolutionised the labour productivity and quality and the individuals' opportunity to benefit from public services and private commerce. Furthermore, the achievements in the management and organisation of records referred to inhabitants, in the public activity, or clientele, in the generally private activity, are extraordinary positive and representing a nowadays inevitable task for

[1] Corresponding author: Vittorio Nicolardi, email: vittorio.nicolardi@uniba.it

statisticians and computer scientists. The production of raw data gathered for a mere administrative purpose represents an essential and precious source of information for scientists and analysts, although many issues related to the same administrative nature of data need to be faced. The problems in using administrative data sources are various and involve many aspects of the modern scientific challenge of profiting from their enormous potentialities.

First of all, it is crucial to bear in mind that administrative data are a little part of a more complex phenomenon occurring over the last two decades, and nowadays known as Big Data. As Japec et al. (2015) argue, the popular term Big Data is a generic way to describe a rich and complicated set of characteristics, ethical issues, practices, analytical techniques and outcomes associated with a large set of data. The generic term "Big Data" refers to the large volume of information and the data variety and velocity, how data are created, the analytical process to analyse them and make inference from them. Big Data are, very often, secondary data, which means that they are collected for another primary use, typically not statistical, and consequentially show an unstructured nature of the information. They can derive from various data sources that can include structured, semi-structured and unstructured data and very often be independent of each other (Pusala et al., 2016). Big Data was historically derived from the physical sciences, where the development of new analytical techniques alongside advancements in new technologies and research laboratory instruments allowed scientists to achieve extraordinary results. Without being poor in the exemplification, the CERN's expertise in Big Data is the most immediate and valuable citation in this sense. Handling and treating large data sets is also valuable in the analysis of biological and clinical phenomena. In fact, in the scientific debate, it is well recognised how the biomedical Big Data framework has opened new opportunities to enhance the understanding of disease heterogeneity in humans (Hamada et al., 2017) and the benefits and challenges that the Big Data approaches of analysis experience in the cell and molecular biology (Dolinski and Troyanskaya, 2015).

In the economics and social sciences, Big Data information capacities are a new opportunity for research. Actually, in most cases, survey data remain the most valid and reliable source of information, although their limits in depicting a society in continuous and fast development are already recognised and discussed in the international scientific context. The role that the administrative datasets assume in the debate over the importance of the official statistics and their necessity of upgrading to the new scientific challenge is internationally significant and well recognised. Although huge administrative private/public databases may not hold all the characteristics we mentioned above that commonly describe the Big Data (e.g.

velocity, variety and volume), they are considered part of them. The definition of Big Data is still nowadays imprecise, and many other characteristics have to be added to complete their meaning and distinguish their different typologies (Harford, 2014). Kitchin (2014) argues that some huge databases may hold even a single Big Data characteristic or a different set of characteristics, but they can still be considered in the Big Data framework. In this sense, the administrative data are undoubtedly more structured and well-defined than other Big Data sources, but the questions of dealing with them, mainly when they are massive sets of data, are still far to be resolved.

The crucial issues related to utilising the administrative data to integrate the official statistical information are widely discussed in the literature (Nordbotten, 2010; Kitchin, 2015; Thomsen and Holmoy, 1998). The main issue is rooted in the primary purpose by which statistical and administrative data are acquired. The statistical data are primarily collected to yield statistics, and this is generally their sole objective; they usually are obtained through surveys or census, and therefore they are expensive in terms of both time and money. As a consequence of their nature, statistical data are created for research, and therefore they are controlled and statistically correct. Statistical data involve almost all variables of interest in the analysed phenomenon and, therefore, they are complete referring to the provided information. Statistical data are yielded through a research plan, and, therefore, informed consent policies and consequentially ethical concerns surround their creation. Although the significance of the statistical information is well established, the National Statistical Institutes (NSIs, hereafter) need other data to compile official statistics that can reflect the lively mutations of the nowadays society. Therefore, data collected and maintained by other nonstatistical organisations represent useful supplementary sources where all the obstacles to their use are resolved, and administrative data are the most valid (Hassani et al., 2014). The administrative data are primarily collected for nonstatistical purposes, but they can yield statistics though they are not created for research; they are usually produced from the administrative processes. Therefore, they are affordable because their collection is not time and money consuming.

As a consequence of their nature, administrative data can be affected by material and human errors, such as duplication, missing values and erroneous information, and therefore they are not controlled and statistically correct. Furthermore, they can frequently derive from various data sources that can very often be independent of each other. The latter is, for instance, the case of information provided by the Public Administration (PA, hereafter): the entirety of the administrative data on the analysed phenomenon frequently requires the

merging of two or more databases that belong to two or more PA offices. The administrative data might generally consider only a few variables related to the analysed phenomenon, and therefore they provide incomplete information. Since the administrative data are not created for research, both statistical concerns and informed consent policies might not surround their production.

Therefore, the enormous potentialities of nonstatistical data differently collected, both voluntary and compulsory, both acquiescing and unintentional, can represent the milestone in the analysis of most socioeconomic phenomena, but necessarily many obstacles need to be solved. Over the last decade, the scientific debate is focused on the resolution of both statistical and juridical issues related to individual privacy (Reiter, 2012; Karr and Reiter, 2016; Wachter and Mittelstadt, 2019), data veracity and possible integration (Di Consiglio and Falorsi, 2015; Calzaroni, 2008), overall assessment of the quality of administrative data sources (Ossen et al., 2011; Daas et al., 2008), data mining and machine learning for the official statistics (Hassani et al., 2014).

In this paper, without entering the technical statistical scenario of the discussion, we provide a first and original successful attempt to align the administrative and statistical data referred to an economic phenomenon that, in Italy and many European countries, still suffers the consequences of the dearth of a complete and harmonised data warehouse. The analysis focused on the Italian real estate phenomenon through an experiment on a sole city. It investigates how the administrative data are powerful in adding new information on the phenomenon in terms of both volume and value compared with the limited evidence that generally arises from the official statistics yielded by NSIs. The real estate economy has been extensively analysed from innumerable points of view, but all data, which since now were used in literature, are complete concerning the database used but partial on the overall view of the phenomenon. We created a unique administrative database perfectly aligned with the Italian NSI Census database starting from huge independent databases managed by autonomous Italian PA offices. Big Data analytic practice and GIS processes were fundamental to guarantee the exact matching of data and depict the real estate territorial framework in detail. Therefore, as many European NSIs already proceed with their national population census, we used valuable administrative governmental datasets and solved all the Big Data issues involved in the study. Furthermore, as well recognised in the Big Data literature, the availability of a huge amount of data, when appropriately summarised into a comprehensible format, can help private organisations in the decision-making process (Laha, 2016; Olszak, 2016) and be also appreciable for PA. Therefore, as a counterpart of the creation of the full information harmonised real estate

database, we yielded an economic indicator, conveniently graphically referred, to provide policymakers and business managers with an instrumental measure to affect the real estate market and the public fiscal policies.

The paper is organised as follows: the description of the dataset contents and the method designed and employed to create the complete information harmonised real estate database are reported in Section 2; Section 3 describes the high potentialities of the harmonised database through an economic evaluation of the real estate phenomenon; some concluding remarks are reported in Section 4.

2. THE DATA AND THE METHOD

The main scientific obstacle encountered when it is necessary to use integrated and complete information on some socioeconomic phenomenon to analyse its development, trend and correlations is related to all the problems involving the merging of official statistical databases, as yielded by NSIs, and the administrative or differently public databases. The assessment of both the typology of data and their corresponding quality is part of a set of problems usually encountered when it is necessary to work with administrative databases. The typology is fundamental to plan the type of the analysis, while the quality is likewise important to guarantee the reliability of the outcomes. Both issues require great attention when the size of the databases is significantly big because the opportunities relying on the major availability of information risk to be a weakness for the purposes of the studies. Therefore, in this sense, when administrative datasets involved in a study are fairly big to be considered in the Data Science scenario, one of the most delicate phases regards their cleaning and management. In our work, we decided to independently work on each database to pre-process data and select the key features of each database to proceed with the merging action.

2.1 THE DATA

The analysis is restricted to the territory of the city of Bari, in southern Italy, because that is part of a national research project[2] that has also guaranteed data accessibility without incurring any juridical violation because of the privacy issue involved. However, the method developed in this work can be perfectly replicated for any dimensional geographical area when the data provider guarantees the accessibility of databases.

---

[2] The Italian National Project reference is named: Metropolitan cities: territorial economic strategies, financial constraints and circular rehabilitation.

The datasets used to create the full information complete real estate administrative database are fundamentally 7. Six of 7 are independent administrative datasets, generally managed by independent PA offices, and provide autonomous real estate information. Dataset 7 is the official Italian NSI Census Section dataset and the sole statistical data source used in this work. The Italian NSI Census Section dataset has assumed the role of an instrumental dataset to validate the obtained outcomes because it contains, among all data, the official statistical geo-localisation information that is used to yield all the official territorial statistics.

Four of 6 administrative databases belong to the Real Estate Registry (RER, hereafter) and contain all information related to the real estates. Although the PA office is the same, the four databases are independent and autonomous in providing the corresponding information. The last two administrative databases belong to the Real Estate Italian Observatory (REIO, hereafter) of the Italian Revenue Agency (IRA, hereinafter).

The RER administrative databases have been fundamental to create the final harmonised database. The main database is named Real Estate Units (REU, hereafter) and includes a list of records that contain all technical and economic cadastral[3] information of each unit. REU provides items of valuable information referred to the real estate cadastral category to identify the various typologies of units such as, for instance, dwelling or shop or office, the cadastral income and the size of each unit. The size of the database is 283,217 records, without duplication, referring to all units, but 20,240 records lack cadastral income because they belong to units of the F cadastral category that is without income. The other 2 RER databases are functional to build the final database. The first is named Cadastral Identifiers (CI, hereafter) and comprises other cadastral information, mainly the Urban Section and the Cadastral Sheet, Subordinate and Parcel. The CI dataset includes 421,324 records, a number much higher than REU records because of duplication, caused by some administrative change, and the presence of some real estate unit whose record has not been deleted though the building was demolished and is not anymore existing. The second RER database is named Cadastral Addresses (CA, hereafter) and comprises the toponyms of each real estate unit. Toponyms are essential to identify the exact localisation of each unit on the urban

[3] Cadastre is a term that has a precise meaning in Italy but can be misleading when referred to other countries. Italian Cadastre is the official administrative registry where all real estate and land documents, maps and information are reported. For instance, the cadastral income is the real estate economic value that is typically used to calculate the council tax, or the cadastral category is the real estate typology such as, for instance, dwelling or shop or office.

territory. The size of the CA dataset is 668,302 records, and, likewise the previous database, many duplications occur because of the modifications of some toponym and/or building number. Finally, the last RER database includes all the geo-localisation cadastral data and, as the Italian NSI Census Section database, assumes an instrumental role in our work.

The primary REIO dataset of IRA is one of the most reliable sources of data to analyse the real estate monetary value dynamics in the Italian real estate market. The REIO real estate value data are calculated based on the trade price per square meter of the properties. They are open data on a biannual basis referred to the minimum and maximum price for all the different types of real estates at the level of the council territory. To use a univocal REIO value in the analysis, the midrange value for each record was calculated. In the REIO dataset, the council territory is split into homogeneous areas that experience the same economic and socio-environmental characteristics, i.e. the REIO zones. All the zones of the same city are then grouped into five territorial districts that delineate precise geographical portions of the urban space: Centre, Near-centre, Outskirts, Suburbs and Extra-Urban. In this work, the REIO dataset of the city of Bari for the year 2018 was used. The biannual data are referred to the 14 typologies of real estates in Bari for 7,814 records that are already suitable for the statistical analyses. The second REIO database includes the geo-localisation data of the REIO zones and, as the previous geo-databases, has been instrumental to the creation of the final harmonised database.

2.2 THE METHOD

The method we have designed and employed is a step-by-step approach that starts dealing with the issues of the databases in their high dimension and proceeds to create the full information harmonised database, i.e. the objective of our work.

As we have seen in the previous section, the three fundamental RER databases experience incomplete or erroneous records, duplication and missing data. Therefore, without affecting the statistical significance of the analysis, first, we have deleted the incomplete records in the REU database to guarantee a homogeneous dataset in terms of information. Afterwards, a cleaning action has also involved the CI and CA datasets because of their duplicated records. In particular, the cleaning of the CI and CA databases was carried out using two distinct fields, respectively the Protocol Number field and the Sequential field, that report the several modifications that involved the real estates over time. Finally, we obtained 3 RER databases that are numerically homogeneous in their size. Table 1 shows the size of all databases we used in this work in terms of fields and records, original size and final size after cleaning when occurred.

Tab. 1: Database contents

Database	Original Size		Final Size	
	Fields	**Records**	**Fields**	**Records**
Real Estate Units	29	283,217	9	262,977
Cadastral Identifiers	13	421,324	6	262,977
Cadastral Addresses	5	668,302	5	262,977
Cadastral GIS	27	83,092	4	82,576
Real Estate Italian Observatory	24	7,814	6	7,814
REIO GIS	14	35	14	35
Italian NSI Census Sections	4	82,576	4	82,576
Final Harmonized Database	-	-	26	262,977

Once the numeric and structural homogeneity of the RER database has been obtained, the successive step is to merge them. The Cadastral Office Real Estate Identification Code (COREIC, hereinafter) has been identified as the only plausible merging field because that is the sole in common between the three databases. Therefore, the Thorough Real Estate Registry (TRER, hereafter) database was created utilizing COREIC.

The final step of our approach is the creation of the omni-comprehensive database that includes all the real estate cadastral and monetary information. Therefore, TRER and REIO databases have be perfectly aligned to obtain the full information harmonised database of the real estate phenomenon. In this sense, it is worth noting that they are unlinked and not directly connectable through any field, although they are referred to the same object. Therefore, solving two technical problems involving the procedure to merge the two databases becomes a crucial issue.

The first and most crucial issue is related to the specific territorial context differently defined in each database: in TRER, the territorial context is the single Cadastral Parcel, while in REIO, the geo-context is the REIO zone. GIS techniques have been very precious to surmount the obstacle in this phase. A GIS procedure based on the three instrumental databases previously described, i.e. the Italian NSI, the RER and the REIO geo-localisation databases, has been created to align the different geo-contexts of the fundamental TRER and REIO databases. In particular, the Italian NSI database has helped to precisely locate all TRER data per each real estate unit within the census sections through the second instrumental database that belongs to the Italian RER. The REIO instrumental database provides the geo-localisation of the REIO zones. Therefore, the GIS procedure has yielded two distinctive maps that we have overlapped to obtain a new database that we decided to name BRIDGEDB. BRIDGEDB connects the REIO zones with the Italian NSI

census sections. In other words, the procedure allows to link the Italian NSI census sections, and indirectly all cadastral units, with each REIO zone and, consequentially, the statistical information with all the administrative data.

The second problem in the alignment of TRER and REIO is related to the real estate typologies because they are differently classified in the two databases. We have, therefore, built a Code Conversion Matrix to relate the two different classifications and surmount this second obstacle. The simultaneous use of the Code Conversion Matrix and BRIDGEDB has finally allowed us to assign the REIO real estate midrange value to each real estate unit for each cadastral category and compute the market value through the cadastral size of each real estate unit. Therefore, the final database includes all the harmonised cadastral and market data for each real estate unit and is entirely consistent with the official statistical information.

3. AN ECONOMIC HARMONISED EVALUATION OF THE REAL ESTATE PHENOMENON

The full information Harmonised Real Estate Database we created allows us to analyse many aspects of the real estate phenomenon that have always been considered as known but never properly quantified. As a counterpart of creating the database, we have decided to yield an economic indicator, conveniently graphically referred to, that can be a useful instrumental measure for policymakers and business managers. Therefore, the monetary differentials between the cadastral incomes and the real estate market monetary values have been calculated to highlight the distance between the two values. The analysis has involved the whole territory of the city of Bari and is referred to only three typologies of real estate units: the Civil Dwellings, the Economic Dwellings and Villas and Detached Houses. The Figures that follow depict the percentage differentials with respect to the cadastral values for each typology. The darkest areas experience negative percentage relative differentials, which means that the cadastral values are significantly lower than the market values. The opposite is the case when the market values are lower than the cadastral values, and the colours are lighter.

The analysis of the relative differentials highlights differentiated values depending on the city neighbourhood and dwelling typology, and their graphical references show an interesting distribution on the territory. Figure 1 describes the percentage relative differentials for the Civil Dwellings in the city of Bari for the year 2018. As we can see, most of the cadastral values are significantly lower than the market values, up to 80% in some cases, and it is mainly located in the centre or near-centre city areas where some rehabilitation action was made, and some

Fig. 1: Percentage relative differentials. City of Bari Civil Dwelling. Year 2018

others are still on progress. The opposite is the case for the Economic Dwellings. Figure 2 shows the percentage relative differentials for the Economic Dwellings in the city of Bari for the year 2018. On the map, it can be noted that the slightly grey areas are more predominant than the darkest. Differentials are, on average, lower in value than the Civil Dwellings case and in many city areas positive, namely, the market values are lower than the cadastral values. The economic evaluation of

Fig. 2: Percentage relative differentials. City of Bari Economic Dwelling. Year 2018

the last outcome is very important if we consider that the areas involved are outskirts and suburbs, namely those neighbourhoods of the city that are more peripheric and slightly economically depressed and paradoxically, the incidence of the local fiscal policy is relatively higher than in the centre of the city. It is noteworthy that an area adjacent to the city centre (near the commercial port) shows very different differentials than those of the adjacent neighbourhood.

The last case of the analysis is depicted in Figure 3 that shows the percentage relative differentials for Villas and Detached Houses in the city of Bari for the year 2018. As we can see on the map, except for the touristic area in the South-East of the city and the residential area adjacent to the airport in the North-Est, the rest of the city shows a critical situation where previously residential areas today are economically depressed. Villas and Detached Houses have fallen in their market value over time, and today, although their commercial value is much lower than the cadastral value, they are still fiscally considered luxury houses.

Fig. 3: Percentage relative differentials. City of Bari Villas and Detached Houses. Year 2018

4. CONCLUDING REMARKS

The nonstatistical data collected in different ways represent the very innovative advancement in analysing most socioeconomic phenomena and although their role is internationally well recognised and debated, they are still far from being fully integrated into all the official statistics. Over the last two decades, all the NSIs are

working to deal with all issues related to the new scientific challenge that also involves the data science context and can revolutionise the official statistical production, but much work still needs to be done. In particular, the role of the administrative data is under investigation and, although it is worldwide scientifically recognised its importance and validity, not all the questions related to this data official use are resolved.

The analysis we provided in this work is unique and original in its attempt to describe the phenomenon of the real estate economy that still suffers the consequences of the dearth of complete information because of the practical nonexistence of a full information harmonised database that includes all data involved. Starting from 4 administrative databases, which we defined "fundamental", and 3 additional administrative databases, which we defined "instrumental", all belonging to autonomous PA offices, we have created a complete Harmonised Real Estate Database.

The extraordinary potentialities of the information Harmonized Real Estate Database are huge and involve many aspects of the PA activities on the one hand and the household/private business on the other. The alignment of cadastral incomes and market values we yielded in this work is without precedent in the literature. The alignment can refer to entire city areas or portions of neighbourhoods or even single units, and such detailed data are the first significant outcome in this sense, though that is only referred to the city of Bari, in South Italy. The empirical evaluation of the real estate phenomenon in the city of Bari we yielded significantly quantifies the magnitude of the differentials, precisely geo-localised, between the cadastral values and market values. Considering that the two values are fundamental for respectively the local fiscal policy and the real estate market, the strength of the outcome is intuitive. It is, however, important to highlight that the methods we have designed and employed can be replicated for any geographical area.

Furthermore, we have perfectly merged the real estate administrative data with the official statistical data. One of the instrumental databases we used is the Italian NSI Census Section database, which has been important to validate our outcomes. Therefore, our Harmonised Real Estate Database can be perfectly connected with all official Italian census data, and the opportunities for new socioeconomic analyses are vast.

REFERENCES

Calzaroni, M. (2008). Le fonti amministrative nei processi e nei prodotti della statistica ufficiale. In *Atti della Nona Conferenza Nazionale di Statistica.*

Daas, P., Arends-Tóoth, J., Schouten, B. and Kuijvenhoven, L. (2008). Quality framework for the evaluation of administrative data. In *Proceedings of Q2008 European Conference on Quality in Official Statistics.* Statistics Italy and Eurostat, Rome.

Di Consiglio, L. and Falorsi, D. (2015). Different contexts for the statistical use of administrative data. In *Proceedings of Statistics Canada Symposium 2014 on Beyond traditional survey taking: adapting to a changing world.*

Dolinski, K. and Troyanskaya, O. (2015). Implications of big data for cell biology. molecular biology of the cell. In *MBoC.* 26(14): 2575-2578.

Hamada, T., Keum, N., Nishihara, R. and Ogino, S. (2017). Molecular pathological epidemiology: new developing frontiers of big data science to study etiologies and pathogenesis. In *J Gastroenterol,* Volume 52, Issue 3. 52(3): 265-275.

Harford, T. (2014). Big data: a big mistake? In *Significance* vol. 11 issue 5. Pp. 14-19.

Hassani, H., Saporta, G., and S.E., S. (2014). Data mining and official statistics: The past, the present and the future. In *Big Data.* 2: 1?10. Doi: http://dx.doi.org/10.1089/.

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., Lane, J., O'Neil, C., and Abe, U. (2015). On square and ordinal contingency tables: a comparison of social class and income mobility for the same individuals. In *Public Opinion Quarterly,* Volume 79, Issue 4. 79(4): 839-880.

Karr, A. and Reiter, J. (2016). Using statistics to protect privacy. In *S.B. Julia Lane Victoria Stodden and H. Nissenbaum, eds., Privacy, Big Data, and the Public Good: Frameworks for Engagement.* New York: Cambridge University Press, New York: Cambridge University Press.

Kitchin, R. (2014). Data, new epistemologies and paradigm shift. In Big Data & Society. 1. 20539517145228481.

Kitchin, R. (2015). The opportunities, challenges and risks of bigdata for official statistics. In *Statistical Journal of the IAOS.* 31(3): 471-481.

Laha, A. (2016). Statistical challenges with big data in management science. In *R.S. Pyne S. Rao B.,* ed., Big Data Analytics. Springer, New Delhi.

Nordbotten, S. (2010). The use of administrative data in official statistics - past, present, and future - with special reference to the nordic countries. In *Journal of official statistics.* 205-223.

Olszak, C. (2016). Toward better understanding and use of business intelligence in organizations. In *Information Systems Management.* 33(2): 105-123.

Ossen, S., Daas, P., and Tennekes, M. (2011). Overall assessment of the quality of administrative data sources. In *Proceedings of the ISI 58th World Statistical Congress.* International Statistical Institute, Dublin.

Pusala, M., Amini, M., S., Katukuri, J., Xie, Y., and Raghavan, V. (2016). Massive data analysis: Tasks, tools, applications, and challenges. In *R.S.e. Pyne S.*

Rao B., ed., *Big Data Analytics. Springer,* New Delhi, Berlin-Heidelberg: 115-162.

Reiter, J. (2012). Statistical approaches to protecting confidentiality for microdata and their effects on the quality of statistical inferences. In *Public Opin Q.* 76(1): 163-181.

Thomsen, I. and Holmoy, A. (1998). Combining data from surveys and administrative record systems. the norwegian experience. In *International Statistical Review.* 66(2): 201-221.

Wachter, S. and Mittelstadt, B. (2019). A right to reasonable inferences: Rethinking data protection law in the age of big data and ai. In *Columbia Business Law Review*.