European Commission

JRC TECHNICAL REPORT

Multipurpose synthetic population for policy applications

Hradec, J., Craglia, M., Di Leo, M., De Nigris, S., Ostlaender, N., Nicholson, N.

2022

Joint Research Centre

EUR 31116 EN

Contact information
Name: European Commission, Joint Research Centre (JRC), Digital Economy Unit
Address: Via Enrico Fermi 2749, 21027 Ispra (VA), Italy
Email: Jiri.HRADEC@ec.europa.eu

EU Science Hub
https://ec.europa.eu/jrc

JRC128595

EUR 31116 EN

PDF ISBN 978-92-76-53478-5 ISSN 1831-9424 doi:10.2760/50072

Luxembourg: Publications Office of the European Union, 2022

How to cite this report: Hradec, J., Craglia, M., Di Leo, M., De Nigris, S., Ostlaender, N., Nicholson, N., *Multipurpose synthetic population for policy applications*, EUR 31116 EN, Publications Office of the European Union, Luxembourg, 2022, ISBN 978-92-79-76-53478-5, doi:10.2760/50072, JRC128595

Contents

Acknowledgements

This work would have never materialised without people who invested their time and knowledge, and who patiently communicated their needs and use cases to the team. The close collaboration and knowledge sharing between units A.5, I.2, I.3, B.1, E.1 and B.6 made this possible.

Our big thanks belong to Jutta Thielen – Del Pozo, who from her own experience understood and encouraged us to focus on the concept of smaller footprint policy.

We would like to thank the JRC Competence Centre on Modelling for years of guidance and feedback.

We are grateful to Kathleen Zoonnenkindt of Housing Europe who has provided invaluable feedback to the Amsterdam pilot, and patiently and iteratively taught us policy design thinking, co-design and many social aspects of community and archetype definition.

Sanne Hettiga of Geodan NL has provided us not only data and consultations for the Amsterdam pilot but also real-life experience and boundary conditions that made our models much more realistic.

We would like to acknowledge the amazing work of the French Office for Statistics (INSEE). The more time we have spent working on the rolling census data, the more we had to admire the vision of their publishing. Also, feedback has been very quick whenever anything was needed.

The Dutch Office for Statistics (CBS) gave us access to complete microdata for verification of our models based on disaggregated statistics. CBS promptly provided step-by-step guidance, technical and methodological support whenever we needed.

Francesco Giusti spent more than half of 2021 exporting and compiling cancer patient records and guiding us through intricacies of health data encoding, privacy protection, data consistence and application use cases.

We would like to thank especially Prof. Henk Scholten, Steeven Luijtens and Prof. Michael Blakemore for their counselling and critical inputs to our work. Their formulation of policy needs, critical review of the proposed policy process and alternatives helped making this report much more scientifically sound and policy relevant.

Authors

Jiri Hradec, Massimo Craglia, Margherita Di Leo, Sarah De Nigris, Nicole Ostlaender, Nicholas Nicholson

Abstract

The purpose of this report is to study prospective advantages of using individual-level privacy-unburdened synthetic data in various types of policy advice applications.

The level of granularity of data used in policy making support has been a result of a painful choice between individual but complex and often inaccessible data, and the data that were aggregated to protect and consolidate the individual information. Higher data aggregations are more convenient to use but tend to hide details, and in effect may lead to policy bias and cumulative disadvantages of population groups.

Thus the key advantage of using the individual data is not the level of detail but the level of structure. Data aggregates annihilate important knowledge from data and lead to skewed policies. If 30% of population in the study area have university education and 30% are over the age of 65, how can we tell if any, some or all people over 65 have university education?

Data synthesis is a decades old concept used from imputation of missing data to generating artificial populations. Since 2019 the new generative adversarial network methods are emerging with promise to create realistic synthetic replicas of data. Such a data retain most of the statistical qualities and information richness of the original data yet fully protect the privacy. Unlike older privacy protection methods depending on cutting away privacy-disclosing information, the synthesis methods generate data from their original distributions. Since data are generated, there is no trackable record-to-record similarity.

The report demonstrates three levels of the data synthesis and their applications: The first use case describes “poor man” statistical data disaggregation in the Amsterdam policy instrument study. In the second use case the structured census-borne statistical personas we applied in activity-based modelling. Deep-learning based synthesis of hierarchical cancer patient records was studied in the third use case together with validation methods.

Unhindered wide access to detailed (population) data has a potential to give rise to new policy instruments. Even more importantly, such access can enable full public and academic scrutiny of these instruments since all data used in decision making could be shared.

Executive summary

This report reflects on three years of learning experience in identification of the best demographic data to enable the inception of better-targeted policies with smaller regulatory footprint. While using population microdata directly would be the optimal solution, the obstacles, such as the risk posed by re-identification of citizens, for their use are insurmountable and had to be compensated by better alternatives. This report explores pros and cons using synthetic population generated via statistical methods or deep learning.

There are key barriers to deployment of synthetic data at scale: access to computational capacity, how and why to choose a synthesis method, how to measure the results, and often, how to preserve ownership of the data. Thus, synthetic methods may be deliberately avoided in spite of the overwhelming policy opportunities and reliability of advice based on synthetic data

However, synthetic population models facilitate the application of novel methods for data-driven policy formulation and evaluation. This report showcases three applications of structured population such as population activity-based modelling, knock-on effects of selective lock-downs during the COVID-19 pandemic, investigative analysis of existing policy instrument design in the energy transition domain, and applications for synthetic cancer patient records.

Furthermore, we delve – from a methodological standpoint - into how a synthetic representation of the real world data preserves the statistical qualities of the original dataset with no record-to-record similarity and very little degradation of the information. At the same time, the data once generated do not need any infrastructure for consequential processing as they become just a file that can be stored and shared.

Indeed, the primary aim for data synthesis is to allow external parties to access and use the synthetic data and to have a high degree of confidence that the inferences made on the synthetic dataset is meaningful also with respect to the real dataset. This also means that at a reasonable level of aggregation, the results from analysing the real and the synthetic dataset are indistinguishable.

Correctly performed synthesis introduces controllable and well described distortion of the original data, which is just a small price to pay for the availability of highly granular privacy unburdened data. As described in the chapter on conclusions, such data can become the unifying bridge between policy support and computational models, by unlocking the potential of data hidden in silos; thus becoming the key enabler of artificial intelligence in business and policy applications in Europe.

1 Introduction

Context

Legislation can have many purposes (Kosti, 2019): to regulate, to authorize, to outlaw, to provide (e.g. funds), to sanction, to grant, to declare or to restrict. In other words, to nudge or enforce certain behaviour beneficial to (hopefully) the whole society. Every single policy change creates new winners and losers (Brick, 2018). The key role of politicians is to make sure that satisfaction in society settles to (at least local) optimum[1]. Economic interests, power plays, individual needs and liberties, social cohesion, environmental threats, future needs, cross-border pressures, all claims need to be settled in one big cauldron where the magic potion called functioning society has been brewed. Equilibrium needs to be reached to prevent conflicts. But is the equilibrium reached the best one achievable or is it just the local optimum of the policy function?

There are strong players and there are many weak players in society (Squires, 2005), those who may be silent, excluded, disadvantaged. By focusing on "the average" as our statistics suggests, we flatten the society into a uniform low-entropy non-descriptive macroeconomic mass. One of the European fundamental rights "no one left behind" calls for more complex description of the society than just one number. One of the solutions is the introduction of a more complex policy instrument mix stemming from behaviour response mapping. OECD publications are a standard source of instrument mix frameworks.

The instruments respecting societal, economic and environmental complexities for policy support[2] actually exist and take the form of a policy instrument mix utilizing e.g. sectoral convergence and intersection[3]. By mapping known unknowns, these instruments mix study the under-examined behavioral issues where the real structure of the society will be better captured as distributions and graphs instead of one-figure aggregated singularity.

Until recently, aggregated population data have been the staple of policy analysis and implementation. They deliver consistence and comparability of information on inhabitants and are very instrumental in policy design. Yet this aggregation leads to delayed and distorted image of detail, which gets amplified when different aggregated data are combined. Policy makers' ability to effectively react, respond, mitigate, and improve the situation are largely dependent on their knowledge of the population involved, the socioeconomic situation and several other spatial and economic factors.

1.1 Problem definition

The slow, but verifiable, governmental data flows are coming under heavy scrutiny as the majority of their qualities are fading away in comparison to timely, accurate, relevant, and inherently complex big data. The recent Coronavirus outbreak, similarly to the 2008 economic crisis, has shown the gaps in governments' knowledge when it is necessary to react rapidly to a critical situation. In both cases, the key unknown was the behaviour of society actors: companies in 2008-9, and citizens in 2020-1.

We need behavioural data to create realistic environment for the new legal post-fact environment. Once a domain of random polls and surveys, during early 2000s the big platforms started massive onslaught on the people's privacy. Google and Facebook app (and others too) collects metadata, i.e. behavioural data, at overwhelming rate. Not only they know when the user wakes up, charges their phone, where and how they commute to work, where they work, work intensity, when they read their email, when and how long they communicate with whom, how often they go shopping and for what and what they watch in which cinema. This information has been contextualized: tracking cookies via "Off-Facebook Activity"[4] make available to the platforms' other providers search data, location information, shopping preferences, and every such a hint leads to linking of a specific behaviour to actionable and commercially exploitable linkage[5] that can be bundled and sold as a product[6]. All data stored in a huge network graph[7] enable

[1] https://www.oecd.org/governance/policy-framework-on-sound-public-governance/
[2] https://www.oecd.org/publications/debate-the-issues-complexity-and-policy-making-9789264271531-en.htm
[3] https://journals.sagepub.com/doi/full/10.1177/2158244019900568
[4] https://www.wired.com/story/off-facebook-activity-privacy/
[5] Google knows everything, because we tell it everything https://www.theceomagazine.com/business/innovation-technology/google-data-privacy/
[6] If You're Not Paying For It, You Become The Product https://www.forbes.com/sites/marketshare/2012/03/05/if-youre-not-paying-for-it-you-become-the-product/
[7] https://medium.com/coinmonks/tao-facebooks-distributed-database-for-social-graph-c2b45f5346ea

efficient search for even non-intuitive links between people, clustering people into behavioural groups, extraction of the typical groups of interest.

Advertisement always fuelled the trade, but recent sheer abundance of personal and contextual data for the modern behavioural science led to "substantially deeper insight into human traits and biases that might be (mis-)used in order to influence how people behave or think without even noticing"[8]. Nobody likes being offered advertisement on something they do not like. But subliminal messages tailored to specific user groups locked in their echo chambers can and have influenced voting results[9] and political radicalization (Almagro, 2022).

If the ambition of legislation is to change behaviour (Collins, 2015), and the previous behavioural data have been held by internet platforms and not by the governments, how this disproportional availability of data will reflect on ability to nudge (soft regulate) the people? The time dimension is also very relevant here. To design and conduct a complex survey takes months, and the results have often been reused for years. The platforms just collect on a minute basis whatever they need, or through gamification get real non-fabricated data. And if the study was incorrect, they can easily run it again.

The behaviour of certain sociodemographic segment is crucial, going beyond the information about a specific user. Super precise advertisement targeting is just a surface scratching compared to building daily assistants that can guide a person through their daily obstacles by nudging and offering solutions that worked in the past either for them or the community. Even those staying off the grid can be described if a few data points can be provided, even void has a border good enough to describe those within.

This situation creates a major knowledge imbalance – global internet platforms already powerful enough have information on current and possible future population behavior at level vastly exceeding the information the governments have. So what is going to happen next? Outsourcing the policy making and policy efficiency assessment to the internet platforms because they know the reality better and can easily collect the information?

1.2 Proposed solution

Behavioural profiles require individual data as anchor to be linked to. But how to link behavioural response learnt from models, surveys and big data to individuals while ensuring credibility and privacy?

Using individual data is difficult, complex, legally and methodically challenging and computationally intensive. Our proposal is the shift towards creation of a **synthetic population graph that grasps the statistical qualities of the original population**. Digital twins where the demographics of the synthetic individuals have been fused with real-life and real-time data providing even higher quality and complexity.

In this report, we delve into a range of techniques to generate this synthetic population graph which can be deployed as an additional instrument to policy analysts and policy researchers. Indeed, one could envision its deployment in open, privacy-safe population sandboxes based on such synthetic data as safe modelling environments for rethinking the whole policy framework. Data fusion, synthesis and practical digital twins are just a few examples of massive advancements which could be facilitated by the creation of this common reference – the probabilistic synthetic population.

While many innovative demographic data sources have been successfully introduced to policy making process (Bosco, 2022), from big data to satellite images, they still lack the detail and complex structure such as a person in a household.

On the other hand, the probabilistic nature of synthetic population enables linking behavioural models to population. Furthermore, by using location of the synthetic peoples' home, workplace, school as well as points of interest, there can be a link established even to where this behaviour takes place. There are several large scale information sources on behaviour already available, from EuroBarometer[10] to HETUS[11] or surveys.

The quality and consistence of input data is pivotal for the consistence of computational models used in policy making. While microeconomic simulations depend on microdata samples, every simulation receives a different sample, statistical representative but hardly repeatable. Utilisation of a unified

[8] Nudging and Digital Platforms Workshop: https://www.hiig.de/en/events/workshop-nudging-digital-platforms/
[9] https://www.theatlantic.com/technology/archive/2017/10/what-facebook-did/542502/
[10] https://europa.eu/eurobarometer/
[11] https://ec.europa.eu/eurostat/web/time-use-surveys

Europe-wide structured synthetic population data with well described data degradation will lead to similar precision in inter-model comparison while massive improvement in inter-model consistence.

This report documents research and experiments on several approaches to population data synthesis, ranging from recreation of the population from statistical aggregates to utilisation of upstream rolling census model to deep learning. Every method has its pros and cons (see Table 1), from realism of the population produced to computational intensity and knowledge required.

In Table 1, we summarize the characteristics of such approaches for individual data usage for policy purposes and, in what follows, we shall explore their application in the context of three use cases, detailed below in Section 1.3. We note that other synthesis methods such as differential privacy or fully conditional generation (El Emam, 2020) are outside the scope of this report as they do not present substantial difference in use for policy making and assessment purposes.

Table 1. Statistical methods for modelling population data.

	Microdata samples	Disaggregated statistics	Upstream population model	Deep learning
Strengths	Most realistic	Least realistic, simplistic, very low detail preservation	Highly consistent and realistic, details obfuscated	Rather consistent, realistic, preserves both detail and privacy
Weaknesses	Prohibitive privacy burden	Difficult to control esp. in small populations, does not scale well.	Depends on properties and representativeness of the upstream model, does not scale	High knowledge and computational demands, blackbox
Opportunities	Trusted Computing (TC)[12]	Simple to use, methods known for decades	Rolling census is becoming an interesting statistical product[13]	Creation of continent-sized populations
Threats	Highly prohibitive cost of the TC	Can lead to wrong results due to oversimplification	Minimal	At the moment high-quality synthesis systems are only commercial

1.3 Structure of the report

The report is structured in the following way: in Section 2 we delve into the policy context relevant to the use of synthetic population data, while in Section 3 we concisely present the methods for synthetic data generation and the assessment of their utility and privacy levels. Sections 4-6 build on the previous ones by presenting the use of synthetic population data in three different use cases that are, in turn, linked to different policy actions, Our purpose in presenting such diverse use cases and contexts is to showcase the ductility of synthetic data, which makes it a tool suitable for the support of a very large array of policy needs. Lastly, we take stock of our analyses in Section 7, concluding with a summary of the lessons learned and commenting on the opportunities and challenges offered by synthetic data.

Here below we concisely summarise the three use cases that shall be presented in Sections 4-6.

1.3.1 The French use case – structured population graph for policy advice

The need to move from nudge to behaviourally informed regulation (Alemanno & Spina, 2014) needs to be based on access to behavioural data, obviously, but also to a common data matrix to which the interesting attributes can be linked. Behavioural profiles cannot be linked to aggregated data, as all the valuable information will be lost, or to individual microdata due to privacy reasons.

[12] https://en.wikipedia.org/wiki/Trusted_Computing
[13] https://ec.europa.eu/eurostat/cros/content/rolling-census_en

One of the most realistic population synthesis methods is disaggregation of rolling census data. The rolling census has been run in France every year. Resulting statistical personas (a profile representing roughly 5-500 people) were linked to households, houses, workplaces and schools, and to activity patterns linking all the places together. This complex graph has been successfully applied in studying COVID exit strategies by examination of links between person, their profession, economic sector, travel patterns and household composition, as we described in (De Groeve, 2020).

The same population was used to build an activity-based model to study patterns of travelling in the City of Lille. We have built the city-wide activity-based model to demonstrate the ability of linking behavioural profiles to population-based graph. While everybody behaves somewhat differently than in the behavioural model, the common patterns start emerging from re-aggregation of all people as a group behaviour. This population graph network can be continuously extended and updated with new information.

The probabilistic nature of the virtual individuals can facilitate to better map opinions, emotional landscapes, personal preferences and obtain group behaviour that may explain effects we may not be aware of.

1.3.2 The Amsterdam use case – investigative study of a policy instrument

Predicting population behaviour in reaction to events, be it crisis or a new policy instrument, is an immensely important task and is supposed to lead to policies better fitting the reality. To this end, more complex mapping of real individual behaviour for verification and exploration is needed and modellers need to know how people really behave in their respective group (Ciriolo, 2011; Alemanno et al, 2012) to make their models trustworthy.

Synthetic population addresses this need, by allowing to measure quantitatively the representativeness of samples, cross-checks between agents, location and the researched behaviour. Furthermore, **group behaviour** can be practically injected into synthetic population microdata - static or probabilistic linkages of attributes defining the group. This can be extracted typically from surveys or probabilistic mapping of group attributes to microdata as we have done in the Amsterdam pilot.

In our second use case, we have demonstrated on the Amsterdam energy policy use case why demographics should play a bigger role in policy instrument design and how communities should become the real addressee. Shift from data aggregates to recreated individuals lead us to discover sociodemographic underlying issues relevant for the green energy transition. Generous co-funding scheme falls flat if the receiving party has no interest or means to contribute to the funding provided, especially elderly and people living in energy poverty.

1.3.3 Individual patient records for free download?

In our third use case, we have applied data synthesis on extremely sensitive data, cancer patient records, using deep learning methods. The resulting dataset has shown astonishing level of realism (are we looking at the original or the synthetic data?) while maintaining all the privacy test. Resulting data not only can be shared freely, but also can help rebalance under-represented classes in research studies via oversampling, making it the perfect input into machine learning and AI models.

Indeed, deep learning methods have become the state of the art in data synthesis, generating synthetic data that are capable of retaining much more than 95% of the value and information of the original dataset while providing practically 100% privacy in large datasets. This accuracy allows using synthetic data as a replacement for actual, privacy-sensitive data in the actual research and policy analytics. And also, to use the real-like synthetic data to train the algorithms that can be later used in trusted infrastructures storing the real individual microdata.

The synthetic data from deep learning are not, however, a silver bullet to every privacy-related problem. Constraints of minimum data quantity, outlier replication or specificity of the generated data are to be considered and we will delve further in these considerations in Section 3.3.

2 Policy context

Digitranscope was a 3-year research project (2018-2020) of the JRC Centre for Advanced Studies focusing on the governance of digitally transformed human societies. The project aimed to provide a deeper understanding of key aspects of digital transformation to help policy-makers address the challenges facing European society over the next decades (Craglia et al, 2021). One workpackage researched and modelled techniques and applications on how policy anticipation and formulation as well as evaluation can benefit from the world of behavioural big data without breaching privacy. The work on synthetic data in policy applications continues within the JRC project ENABLED with the aim to utilize synthetic data as complement and in some cases substitute other data sources for European data spaces.

The development of the Digitranscope project coincided with an increasing recognition of the importance of Artificial Intelligence (AI) to master the increasing volumes of big data available on a daily basis. The control of AI and of the data underpinning its development are strategic for the future development of society, and the focus of an increasing geopolitical competition. The European Union has identified technological and data sovereignty as key priorities for Europe and developed several policy initiatives to strengthen its regulatory framework and increase its preparedness to address both digital and green deal transformations. These initiatives are addressing both the governance of AI and of the data economy underpinning it.

With respect to AI, the European AI Strategy and Coordinated Plan initiated with the EU Member States in 2018 built on three pillars: ensure technological developments and uptake of AI in the different sectors, prepare citizens for the socio-economic changes brought by AI, and lay out an appropriate ethical and legal framework. This framework was supported by a High-Level Expert Group on AI that brought together representatives from academia, civil society and industry. The recommendations by this group centred on the concept of Trustworthy AI based on seven key ethical principles: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental well-being; and accountability. The EC's efforts towards Trustworthy AI further progressed on April 2021 with the publication of the AI Act, a proposal for a regulation laying down harmonised rules on AI. Rather than on AI techniques per se, the proposal focuses on particular applications and establishes different levels of risks to fundamental rights and safety, from unacceptable risk (prohibited practices) to minimal or no risks. For each risk level, the proposal defines a proportionate set of requirements that AI system must fulfil.

On the data side, the European Strategy for data encourages the creation of several thematic data spaces in which civil society, the public and the commercial sector can share data. The Data Governance Act aims to facilitate voluntary sharing of data by individuals and businesses and harmonises conditions for the use of certain public sector data. A new key initiative is the forthcoming Data Act, which extends the rights of users to access and share data generated by products or services they use, and together with other legislation covering Digital Services and the Digital Market, prevents the abuse of dominant position by large players in ways that harm citizens, business and consumers.

Open Data PSI Directive (EU) 2019/1024) documents the re-use of data, which is associated with important benefits for society, the environment and the economy. These benefits stem from suitability of data for the creation of value-added services, applications and new, high-quality and decent jobs and from the number of potential beneficiaries of such value-added services based on those datasets. The Implementing Regulation on High-value datasets further extends and improves access to data with strategic role in data sharing as shown in its inception impact assessment,

This combination of legal instruments sets the boundaries for the development of AI technologies in a way that supports the values that are at the base of the European Union, namely respect for human dignity, freedom, democracy, equality, the rule of law, and respect for human rights. In doing so, it sets the development of AI in Europe apart from that of other parts of the world where state, military or commercial interests have the leading roles.

Whilst the framework outlined above focuses on the governance *of* AI and data, the digital transformation offers also new opportunities for new forms of policy design, implementation, and assessment i.e. new forms of governance *with* AI and data. The use of digital twins, gaming, simulation, and synthetic data are just at their beginning but promise to change radically the relationships among all the stakeholders in governance of our society. This report focuses on the work started under Digitranscope on the use of synthetic data for policy and provides an important contribution to both policy and practice.

The identification of sociodemographic barriers to policy implementation has not yet become an intrinsic part of the policy and policy instrument anticipation, formulation and assessment process. However, in this sense, synthetic data have the potential to help building completely new policy instruments due to its ability to replicate attributes of individuals, their opinions and sentiment while preserving their privacy, becoming the key input to understanding such sociodemographic barriers. Synthetic population data could support the policymaking process in, potentially, all fields and the use cases we chose to present are meant to convey such diversity of scope, spanning from health (Section 6), COVID-19 re-opening (Section 4) and energy efficiency policies (Section5). For instance, in the case of the latter, the allocation of structured household into houses could enable creating highly realistic activity profiles in domestic energy modelling. Thus, it could support policies on building energy efficiency or sociodemographic capacity for solar panels installations.

3 Data synthesis

3.1 Methods

Population synthesis has been adopted in various fields, such as urban systems evolution (Farooq et al. 2013; Antoni et al. 2017; Delhoum et al. 2020), infrastructure planning (Ye et al. 2009; Jorosz 2013; Arentze et al. 2007), demographic dynamics (Ironmonger et al. 2000; Sajjad et al. 2016; Namazi-Rad et al. 2014), human-environment interactions, planning, policy development (Stevens et al. 2015), land use modelling, just to mention a few.

Disaggregated synthetic population is input to individual-based multi-agent models (Lenormand & Deffuant 2012; Gargiulo et al. 2010; Thiriot & Sevenet 2020).

The method that is chosen for data synthesis depends on the purpose of the study (Kolb et al. 2011) and, notably, on the format and type of the data that Governments make available.

Census population information are normally available at different granularity in space and time depending on the country. Traditional census data are generally 10 (or 5) years apart. More and more countries are adopting an alternative strategy known as rolling census, which allows updating demographic data every year and reduces the burden on the public[14].

Population synthesis methods are usually based on a limited data sample at very high disaggregated level, that represent from 1% to 5% of the population (Farooq et al. 2013), also known as PUMS (Public Use Microdata Sample), or simply microdata. This is an anonymised subset of census data that Statistics Offices put at disposals of researchers, after making sure of removing every location detail and blurring other information that may allow reverse engineering the identity of individuals. Examples of sample-free population synthesis are less frequent (Lenormand & Deffuant 2012; Gargiulo et al. 2010).

In addition to microdata, also cross-classification tables are needed for population synthesis, in general released by the Statistical Offices as well, that present the joint distributions at various levels of details of 1 to 3 attributes, like for example unemployment by level of education by gender. From such tables it is possible to derive conditional probability for a certain co-occurrence of attributes.

Based on the purpose of the study, some methodologies concentrate on reconstructing the characteristics of the individuals, and therefore put harder constraints on preserving the individual's features at the expense of the accuracy in the household composition, whereas for other studies the household composition is of prior interest. Some studies may also focus on the longitudinal development of the population (aging) whereas some others may depict a certain snapshot in time. Depending on the features of interest for the study, some characteristics might have lower weights or simply be discarded in the effort of preserving the consistency among the features that are valued the most in the study. In addition, data collection might be designed having already in mind the specific research task.

In literature, traditional methods for generating a synthetic population tackle the generation as a fitting problem. Two main families of techniques are the **Synthetic Reconstruction (SR)** and the **Combinatorial Optimization (CO)** (Voas & Williamson 2001; Farooq et al. 2013; Namazi-Rad et al. 2014).

Starting from the microdata (also called "seed"), both SR and CO methods reconstruct the missing records (individuals) using the marginal probabilities as constraints, making sure that the statistical figures of the population are reflected in the modelled population, within some level of accuracy. It is normally necessary to operate a selection of the features that are important to the problem under investigation, neglecting or relaxing the constraints on the remaining attributes (Namazi-Rad et al. 2014).

Within Combinatorial Optimization methods, the **Hill Climbing (HC)** starts proposing a random solution and iteratively tries to improve it, maximizing an objective function, measuring the performance at each loop. Hilltop is reached when the errors are lower than a certain predefined threshold (Namazi-Rad et al. 2014).

Among the most popular traditional SR methods are the **Iterative Proportional Fitting (IPF)**, (Deming & Stephan 1940; Beckam et al. 1996), and the **Iterative Proportional Updating (IPU**) (Ye et al. 2009) techniques.

[14] https://ec.europa.eu/eurostat/cros/content/rolling-census_en

The trade-off in IPF and similar techniques is between the number of characteristics that one wants to model and the computational time needed to take into account all the combinations among them. Therefore, it is usually necessary to decide a priori what are the variables that are most important to the study and give them a higher weight, so that the optimization proceeds at the expenses of variables considered less important. Furthermore, as the IPF methodology entails the estimation of household and individual level joint distributions, Ye et al. 2009 point out that, because household and individual weights will never match, it is necessary to choose beforehand whether to perform the proliferation at the person level or at the household level. This implies that it is impossible to generate a population that fits to more than one purpose. Another limitation of IPF is that a method to measure the quality of the final estimate is currently not available. Furthermore, it is not suitable for sparse population (Tanton, 2018).

Methods in literature are generally modelling either at the household level or at the individual level, neglecting, to a certain extent, the other level. In general, techniques for modelling at household level are based on the Poisson distribution or some modified version of it (Ironmonger et al. 2000; Jorosz 2013).

Namazi-Rad et al. 2014 tried to overcome this dichotomy proposing a nested individuals / households model, whereas Arentze et al. 2009 proposed a method to generate synthetic households based on distributions of individuals. Since the IPF cannot be readily applied if the purpose is to generate synthetic households when the demographic data describes the population in terms of counts of individuals, they propose a two-step IPF procedure.

Also Ye et al. 2009 made an effort to estimate both individuals and households. They introduced the Iterative Proportional Updating (IPU) algorithm. The main difference with IPF is that the latter involves picking all individuals in the households according to the household weights. So the individual's weights are forced to be equal to the corresponding household weights, when in fact they are different. The IPU algorithm, on the contrary, iteratively adjusts and reallocates weights of households during fitting, until households and individuals weights are matched (Kagho et al 2020).

Lenormand & Deffuant 2012 compared two approaches to generate a synthetic population of individuals in households: **sample-free and sample-based (General Iterative Proportional Updating, GPU)** methods. With the sample-free approach, households are built by picking individuals from a pool initially comprising of the whole population and progressively shrinking. This method is less data demanding, compared to the other, but it requires more data pre-processing. This approach requires learning from data the multivariate distributions characterizing the individuals into the households. The sample-free method gives better fit between observed and simulated distribution for both household and individual distribution compared to the GPU approach.

An important problem that affects both SR and CO methods is that they proceed “cloning” the existing records (individuals) rather than synthesize new ones, and this may lead to a loss of heterogeneity compared to the original population. This means that only the combination of features present in the microdata sample will be preserved and further reproduced, neglecting new original combinations that are less frequent or for some reasons were not caught in the sample.

Generative probabilistic models, such as **Hidden Markov models** (Farooq et al. 2013; Saadi et al. 2016) and **Bayesian Networks** (Sun & Erath 2015), allow to generate new combinations of features not present in the original sample.

Bayesian Networks (BNs) are versed in modelling a system as a network of cause-effect interactions, and are particularly suitable for modelling environmental systems, due to their ability to take into account the interplay between various factors in complex systems (Chen & Pollino 2012).

Hidden Markov models are more commonly found in the generation of time series in several fields from economics and finance to physics, pattern recognition, etc.

Farooq et al. 2013 use the **Markov Chain Monte Carlo (MCMC)** simulation technique to draw from the real population distribution. In fact, MCMC allows to simulate a dependent sequence of random draws if the joint distribution of the attributes if the real population is unknown. Unlike IPF, the simulation is able to minimize deviations for both large and small probability values.

Hidden Markov Models are used to infer the underlying state of a certain system for which we can observe emissions. Markovian models are usually used for monitoring the time series of the emission of a certain system, so they make predictions on how the system evolves in time. However, if we consider the vector of features characterising an individual, we can consider the set of possible

instances of a certain feature, as possible emissions of that feature, and each feature can be drawn independently, generating as many instances of individuals as desired, with set of features all different from each other.

With the development of computational power, Deep Neural Networks (DNNs) have seen an unprecedented growth in popularity, and have been incorporated into probabilistic models, creating a new generation of models that exploit deep learning for creating synthetic data: the so-called **Deep Generative Models (DGMs**).

Machine Learning and Deep Learning models are actually game-changers for their ability to learn probability distributions from data using unsupervised techniques and to synthesize high quality data, without needing to know beforehand the scope, or how these data will be used.

Goodfellow, 2016 proposed a thorough taxonomy to classify DGMs with respect to how they represent or approximate the likelihood of the training data.

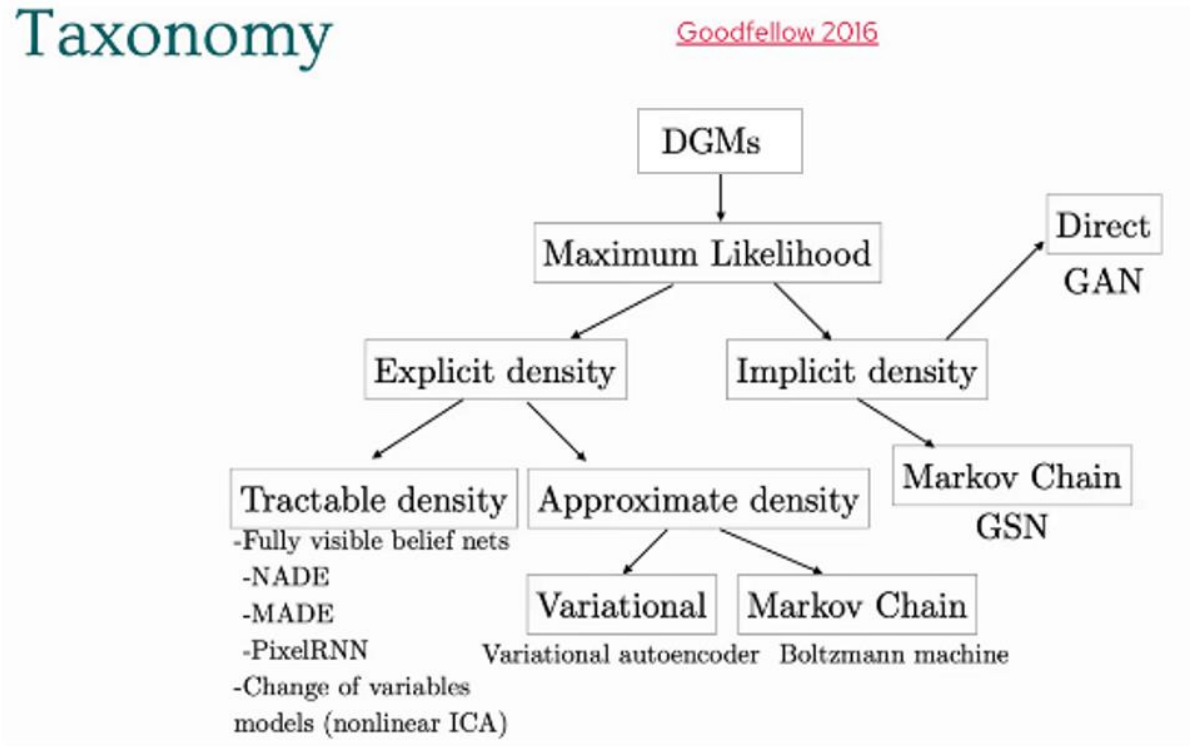


Figure 1 Taxonomy for DGMs classification, proposed by Goodfellow, 2016

We can distinguish two main families of generative models. Models from the left branch make an explicit formulation of the data probability density function (pdf), whereas models of the right branch can generate new data instances without an explicit formulation of the pdf. These two families are also recognized by Girin et al. 2020.

Within the left branch, **Variational AutoEncoders (VAE)** (Rezende et al. 2014; Kingma & Welling 2014) are notable, whereas **Generative Adversarial Networks (GANs**) (Goodfellow 2016; Goodfellow et al. 2014, Goodfellow et al. 2016) belong to the right branch.

The GAN framework stems from the idea of a game between two opposing players: the “Generator” creates the instances of records, as realistic-looking as in its power, and the “Discriminator” evaluates whether each instance belongs to the original (real) population or is a “fake”, generated by the opponent. The Generator improves its performance based on the feedback that receives from the Discriminator. The network converges when the Discriminator is no longer able to tell between a genuine and a fake instance, so it rejects a fake instance with the same probability as it rejects a genuine one. GANs have shown remarkable results in many fields, e.g. the ability to generate photograph-like images, or other types of synthetic data that are indistinguishable from the original data, even by domain experts (Choi et al, 2017).

For their multiple successes in several fields of applications, GANs are considered the go-to for the data synthesis, with the great advantage of not requiring labelled data for training the network (Goodfellow et al. 2014).

Many methods have been developed for data synthesis, some of them are well established, some others are newer and promising, and all of them have their advantages and drawbacks. There is no "one-fits-all method", the choice is always a trade-off between the features that need to be accounted for, and the related computational burden, but probably the main discriminant is the data availability and format.

In a context of full availability of the original data, that is the case of data stewards, the choice of the model to be used for data synthesis depends on the domain of application (use case) and is usually a trade-off between the level of disclosure that can be afforded versus the utility of the data, aka the fidelity between the synthetic and the original data in multiple dimensions, in regard to the use case.

3.2 Use cases of synthetic data

The compound "Data Revolution" was first used in the Report of the High Level Panel (HLP) of the Eminent Persons appointed by the former UN Secretary – General Ban Ki-moon to advise regarding the Post-2015 Development Agenda. The HLP envisaged a Data Revolution as a way to end poverty and promote sustainable development. In recent years, the data demand has seen a dramatic increase from all different sectors, but the data access democratization, key according to the HLP, is still in progress.

It is evident that data availability triggers innovation, and this eventually translates into better services for the citizens. In the Data-driven economy, business is based on the capacity of economic actors to use the amount of information that links the physical world and the digital world, start-ups demand to pitch into the data technology ecosystem for leveraging innovation, from customising services and products to software testing, from technology evaluation to Artificial Intelligence (AI). AI often needs a large amount of data to train models, that can also in turn be used for imputation of missing data or for augmentation of existing data (Synthetically-augmented data sets) to build models more rapidly and efficiently (Kaloskampis et al, 2020).

Researchers need to use and share data to ensure the reproducibility of their claims, as the full independent replication has become a minimum standard for judging scientific claims (Peng, 2011).

In the era of data-driven evidence-informed policymaking, the research behind policy should be as transparent and objective as possible, in order to ensure the trust of the public. Thus, particularly policy-makers should make sure that the data that they process for making informed evidence-based decisions are accessible and open, and the research is reproducible.

In parallel to the data demand, also the awareness that certain data must be kept confidential has surged. Citizens are growing sceptical about releasing personal data, whereas at the same time do not seem particularly aware about the data that every day they volunteer to social networks. Episodes of data breaches have mined the trust of customers in the companies that handle their personal data. In Europe, one of the most stringent data protection laws, the General Data Protection Regulation (GDPR)[15], responds to privacy and security concerns by posing limitations on data processing (collecting, recording, organizing, structuring, storing, using, erasing, etc.) of personal data, onto organizations anywhere, so long as they target data related to people in the EU. It is currently very difficult to share data without concerns of violating privacy.

Synthetic data has proven to be a valid alternative to the original data, whenever this latter cannot be used (Patki et al. 2016), e.g. for enabling data science, reproducible research, official statistics and any other context where privacy is a primary concern (Burgard et al. 2017).

As opposed to real data, synthetic data are de-personalized (not personal) data, and therefore can be used in cases in which the target is not to identify a certain person. Synthetic data is fake data that has the same statistical properties of the real data from which it was generated. In addition, it enables the possibility to enhance certain characteristics that might not be cut by the real data, e.g. outliers, biases, etc. For example, synthetic data can be generated ad hoc if the questions of a survey are correlated to the likelihood of being a survey responder[16].

[15] https://ec.europa.eu/info/law/law-topic/data-protection_en
[16] https://replica-analytics.com/synthesis-tutorials

Beaulieu-Jones et al, 2019 demonstrate a case study in which privacy-preserving generative deep neural networks were successfully employed for sharing clinical data while preserving privacy, incorporating differential privacy into their machine learning models.

Synthetic data can be shared with data scientists and software developers, which can test / develop their software and models with minimal constraints. Once the code is developed, they can send it to a secure environment, suitable for handling personal information, where the analysis is run on real data. Under this model, the data scientists and developers never get access to real data[16].

Another case in which it is useful to use synthetic data is, for example, when the data collection has been designed for a certain scope, for which the respondents of a survey have signed a privacy consent. Afterwards, the same dataset cannot be used for a new (secondary) scope without going back and have the respondents sign the consent again for this latter. In this case, depersonalized synthetic data can be used for secondary scope without privacy breaching[16].

3.3 Assessing the quality of synthetic data: Utility

The quality of synthetic data is highly dependent on the quality of the model that created it. The process of generating synthetic data offers the possibility to set in advance the desired criteria of quality and privacy (Kaloskampis et al, 2020). Once a first instance of data has been produced, one can assess if the criteria are met and, if not, the model used to produce the data can be changed accordingly.

The "utility" of a data set is the suitability of the same to be used for a certain purpose. The data utility of a synthetic data set is maximized when the similarity with the original (real) data set is maximized. Rather than a single measure, utility is normally measured in multiple dimensions.

El Emam, 2020 summarizes seven different strategies that assess different dimensions of the utility of synthetic data (see Table 1). "Not all the methods are supposed to be run on every dataset. The table explains when each utility assessment should be executed. Some types of assessment should be run each time data is synthesized, such as the general utility metrics, the structural similarity tests and the bias and stability assessments. The remaining types of utility evaluation are better used to inform methodology improvements. Rather of being seen as necessary for every dataset that is synthesized, the recommendations for which utility evaluation should be performed are influenced by the costs and benefits of each assessment (El Emam, 2020).

Table 2 Applicability of different utility assessment methods for synthetic data, after El Emam, 2020

Table 1. Applicability of different utility assessment methods for synthetic data.

Utility assessment approach	Explanatory comments	Applicability
Structural similarity	This is critical. If the data is not structurally similar, then that just makes it harder for analysts to use it.	Perform for every data set
General utility metrics	This is critical. Every data set needs to pass a minimal set of utility metrics. This is relatively easy to do because it can be largely automated.	Perform for every data set
Replication of studies	Replication is a convincing way to demonstrate that a synthetic data method can be relied upon. It is a time-consuming process that requires domain expertise.	Evaluate methodology
Subjective assessment by domain experts	This type of assessment of the synthesis methodology can also be quite convincing. It is a more challenging assessment to perform.	Evaluate methodology
Bias and stability assessment	This is a generally useful type of assessment to perform for every synthetic data release. However, the weight of evidence it adds to the utility of a synthetic data set is smaller than the other approaches.	Every data set
Comparison with public aggregate data	When available, comparisons to public data will enhance confidence in a synthesis methodology.	Methodology evaluation
Comparison with other PETs	This type of assessment is useful to perform at some point to help decision makers decide the relative strengths and weaknesses of particular PETs for providing data access.	Methodology evaluation

Raab et al, 2021 recognize two main reasons to evaluate the utility of synthetic data: to compare different synthesis methods for the same data set and to diagnose where the real and synthetic data distributions differ and use this information to improve the data synthesis. They recommend to evaluate the utility

with the propensity Mean Squared Error (pMSE) for the first purpose, whereas, in order to check where the data distributions differ, they recommend to group the original and synthetic data by constructing tables based on their values, and compute measures of difference between the tables.

Liu et al. 2021 use MAPE (Mean Absolute Percentage Error) to measure the divergence between the distributions of attributes in the synthetic population realizations and those from aggregated statistics.

Karr et al, 2006 observe that a highly specific utility measure generally produces a release tailored to a reduced class of analyses, whereas a broad utility measure may generate a release that is “pretty good” for a number of analyses but “really good” for none. The authors distinguish between global and narrow utility measures. Remaining into narrow measures, a measure of utility is the degree of overlap between confidence intervals from the same regression for the real and the synthetic data. When the utility has to be evaluated for multiple applications, an approach is to calculate multidimensional utility measures, one per model.

Woo et al, 2006 present and evaluate four global utility measures that capture differences in the distributions of the original data and the masked data (data that have undergone any statistical disclosure limitations, that can be extended to synthetic data as well) and entail propensity scores, cluster analysis and empirical distribution estimations. The authors also underline the difficulty of interpreting the measures to compare the merits of different masking strategies on an absolute scale. Global utility metrics are not suitable for quantitatively estimating how much utility is lost, and are only one component of data utility estimation. Qualitative considerations can be made on the impacts of the masking on the multivariate statistical distributions.

3.4 The trade-off between Utility and Privacy

When deciding on competing masking strategies, data disseminators should assess the disclosure risk vs data utility of each strategy (Woo et al, 2009; Karr et al, 2006). Preserving the utility of the data sets while providing privacy guarantees is a well-known challenge. There will always be some cases when the demand for individuality (utility) cannot be met without risking on the privacy side (Bellovin et al, 2019).

Differential privacy (Dwork et al, 2011; Machanavajjhala et al, 2017; Wood et al, 2018) is a mathematically rigorous definition that provides strong guarantees on the level of confidentiality that can be achieved by the user, and is used in several contexts. In some cases, however, it may lead to sensible utility reduction, particularly if not especially tailored for the context in which the data set is used. Coupling synthetic data with differential privacy sanitization is a good strategy to achieve a reasonable level of security without compromising the utility too much. The occasional cases in which differential privacy coupled with synthetic data are not enough to preserve security, however, present some good opportunity to improve the privacy model.

Hitaj et al. 2017 for example, presented a situation in which differential privacy is not sufficient to preserve the security, with synthetic data generated under the GAN. The authors demonstrate that any privacy-preserving collaborative deep learning is susceptible to a powerful attack that they devise in their paper. They show that a distributed, federated, or decentralized deep learning approach does not guarantee the protection of the training set during the learning process. Therefore, the authors argue that collaborative machine learning is less desirable than centralized learning. They also show that record-level differential privacy applied to shared parameters of the model is ineffective, and thus recommend applying differential privacy at different granularities.

Huang et al. 2017 created a context-aware privacy framework called Generative Adversarial Privacy (GAP) that allows the data holder to learn privatization schemes from the data itself, leveraging the same GANs technology. The authors demonstrate that the privacy mechanism learned from data in a generative adversarial fashion, reaches the theoretically optimal one. They also claim that their framework can be easily applied in practice, even in the absence of dataset statistics.

Works such as Stadler et al, 2020 that argue that synthetic data do not offer a better privatization scheme respect to classical anonymization of data, are of great importance for improving the methods used for privacy sanitization. Unfortunately, the Authors do not provide evaluation on context-aware privatization methods such as information-theoretic privacy (Hsu et al, 2019; Ding et al, 2021) or the GAP (Generative

Adversial Privacy, Huang et al, 2017). However, they provide a reproducible framework for the evaluation of privacy vs. utility, leaving open future research and application on such methods.

Triastcyn & Faltings, 2019 propose the generation of artificial data that retain the statistical properties of the original dataset as means of granting the privacy of the original dataset. They use generative adversarial networks to draw artificial samples and derive an empirical methodology to assess the risk of disclosure in a differential-privacy-like way.

Hsu et al, 2019 consider the problem of identifying records that may be disclosed without incurring a privacy risk. They assign a “privacy risk score” to each record. This mapping is called the “privacy watchdog” and is based on a record-wise information leakage metric called the information density, or lift privacy. This quantitative measure allows tuning the desired privacy level. Similarly, Ding et al, 2021 proceed on optimizing this scheme.

Platzer and Reuter, 2021 underline that AI-based approaches for generating synthetic data provide a promising novel toolbox in the field of statistical disclosure control (SDC), but, like traditional methodologies, also share the fundamental need to balance data utility vs. disclosure risk.

Reiter et al, 2014 propose a strategy for estimating the disclosure risks for multiply imputed, synthetic data, based on a Bayesian estimation, which can be used as a screening device to identify records potentially at risk and deserving of deeper investigation.

Bindschaedler et al, 2017 generalize the concept of plausible deniability for the application to general data synthesis by establishing it as a privacy criterion in terms of the underlying synthesis probabilities. This framework can be applied to any generative model while keeping the same privacy guarantees defined by the data producer.

The alternative to synthetic data sets is in most cases no sharing individual-level data with the broad public. For census data, the current, most common form of publishing data is releasing univariate, aggregate statistics and summary tables. Microdata are made available to a restricted number of researchers.

Before releasing microdata, data stewards alter them in order to reduce disclosure risks, by removing key identifiers, altering sensitive attributes, swapping some attribute between records, adding some noise to numerical data values and so on. Virtually, all public use data releases have undergone some forms of statistical disclosure limitations (SDL) (Karr et al., 2006). In many cases, data sets are published in the form of aggregated statistics and, for selected users, such as researchers, anonymized microdata.

Microdata are the main input to micromodels and agent-based models. However, the research behind is not always reproducible, if the datasets cannot be shared. Reproducible research should be the main pillar behind evidence-based policy design, as a guarantee of fairness and accountability (Peng, 2011). Access to high-quality individual-based data is in fact foreclosed to the broader community interested in leveraging the technology linked to the data-driven economy.

In conclusion, synthetic data can be used for any use case that do not involve the identification of a person, because there is no one-to-one mapping between synthetic and real people. It is important to underline that the population synthesis is based on models, and as such, models have parameters and a defined domain of applications. The parameters that can be tuned in the model are essentially the fidelity to multiple dimensions of the dataset and the level of privacy, e.g. the granularity of the features representation. A population that has been synthesized for a certain use case might not be suitable for a different one.

Concerning the multidimensional fidelity, for example, some synthetic dataset might need to maintain the general statistics of the original population, whereas another dataset might need the augmentation of certain traits that are not well represented in the original population. All these characteristics can be tuned in the model and need to be decided according to the domain of application.

4 Use case 1. The French population graph synthesis

Our French population generation pilot heavily depended on the enlightened and visionary open data program of the French Statistical Office, INSEE. In-depth analysis, recombination and projections of the data INSEE is publishing has shown absolute consistence and intimate knowledge of publishing data for analytical purposes at a level hard to seek in other countries. The size of statistical archetypes are based on small enough aggregations and allow links to behavioural data, and their weights are balanced to respect location.

Going down this road would be very challenging for many other statistical offices. The time needed for production and verification of a robust population is also a huge challenge to usability.

We have built a synthetic population graph dataset for the whole France generating 63 million synthetic individuals, 22 million households, geo-localised 20 million houses, 10 million workplaces, 5 million schools with relations between them. The probabilistic nature arises from the underlying weighted population model provided by INSEE and data imputation for statistical areas with small populations. Details on the data generation have been included in the Annex A.

The wealth of sociodemographic attributes allows the extraction of patterns and complex enrichment as demonstrated in the use cases. In this chapter, we present two use cases of the synthetic French population.

4.1 Using data effectively to support post lock-down re-opening

The running thread of this report is that data is at the heart of AI applications: being able to access and use effectively the data we have is essential for every company, public organisation or state. We report here as an example the result of work at the JRC to support the European Commission and the member states to assess the relative risks of reopening different economic sectors after the lock-down period. As we have seen, during lock down, which was more or less stringent in different EU countries, only essential services were kept running at all times (e.g. utilities, food production and distribution, pharmaceuticals, essential infrastructures). As the peak of contagion was reached and passed, there was a need to identify which economic sectors to open first to allow the restart of the economy whilst reducing risk of second waves of infection.

To answer this question, several steps were needed:

1) *Create a model of the likely number of daily contacts of each person based on both economic and social activities;*
2) *For the economic activities, identify the relative number of daily contacts of each worker by economic sector, taking also into account for each sector the potential for telework and the proportion of workers commuting daily by public transport in "normal" circumstances;*
3) *Assess the socio-economic impact of risk (by gender and income);*
4) *Assess the spatial distribution of risk, based on socio-economic characteristics of different regions, and commuting patterns.*

For 1), the EU statistics on income and living conditions (EU SILC 2015) data were used to extract information on individuals and households and were combined with cultural participation[17] and enriched by remodelled data on regular use of public transport only available for 2014.

The data were aggregated to obtain personal archetypes (e.g. male, aged 50-59, living in a household of 5 people in highly populated FR3 NUTS area, working as a manager in agricultural workplace with 20-50 co-workers, who has a car in the household and meets relatives daily, while going to cinema and concerts at least once a month). All together 914 520 archetypes were generated representing the whole adult population of France.

[17] https://ec.europa.eu/eurostat/statistics-explained/index.php/EU_statistics_on_income_and_living_conditions_(EU-SILC)_methodology_-_2015_Social/cultural_participation_and_material_deprivation

For each archetype, the number of daily contacts (DC) was calculated under the following assumptions:

- Household size: all people in household meet daily, added to DC
- Apartment house adds 5 (3 for smaller apartment houses) to DC, detached houses add 0
- If schooled: 50 contacts added for all types of school
- Public transport: 50 contacts added to DC if in densely populated area, 30 for medium population, 10 for scarcely populated areas.
- Meeting relatives/friends: 5(10) people added for daily contacts, 5(10)/3 for every week, /10 for several times a month, /30 for once a month, /80 for less than once a month, 0 for no contacts.
- Cinema/concerts/cultural sites/stadium: coefficient 24/265 for going at least three times a year to the event, 2/365 for less than three times a year, zero for every other situation; coefficient was multiplied by 300 for concerts and stadium visits, 50 for cinema and 20 for cultural sites.
- Same approach for voluntary activities, where people on average meet 100 other people on activities related to participation in in/formal voluntary activities and active citizenship.

For 2) and 3) data from the complete EULFS[18] (European Union Labour Force Survey) 2015 database were extracted on professions (isco3d, isco1d codes), economic sector (nace1d, incdecil), by sex/gender, and country. EULFS is a large household sample survey providing quarterly results on labour participation of people aged 15 and over as well as on people outside the labour force and it currently contains data for all Member States, as well as data for Iceland, Norway, Switzerland and the United Kingdom. In total, data covering 218 million economically active people (those with ISCO and NACE codes) were used further in the study. Data on physical proximity by economic sector, and therefore daily contact potential, was estimated based on Dingel and Neiman study (Dingel & Neiman, 2020). The graph below shows as an example the physical proximity (scoring on X axis, the higher the closer proximity to others) vs. income deciles (1-10 on Y axis, 10 is the highest income decile) by economic sector, profession, and gender (blue ring around blob = men, red= women).

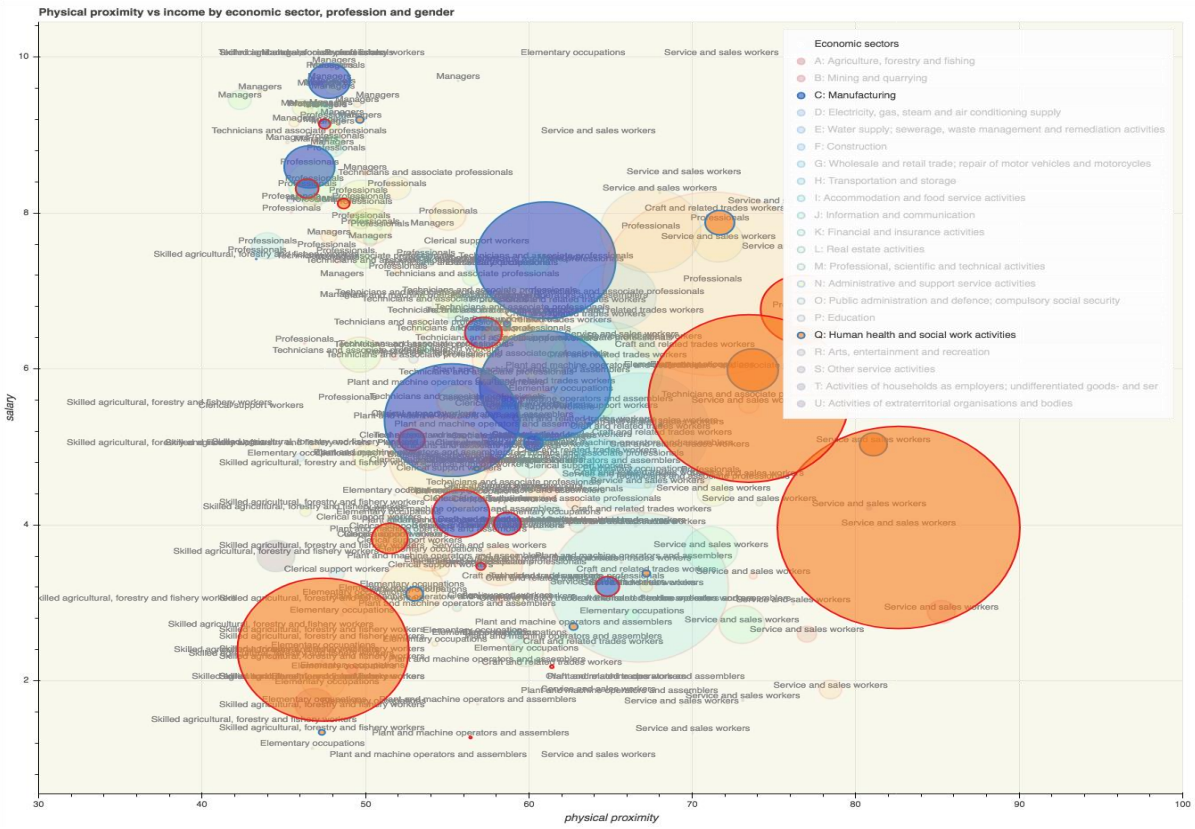


Figure 2 Physical proximity vs income for all sectors with Manufacturing and Health highlighted

[18] https://ec.europa.eu/eurostat/web/microdata/european-union-labour-force-survey

To address 4) we leveraged the synthetic population we have developed. To understand mobility, we have built a model of the commuting patterns of 26 million synthetic French commuters in 2016 with blue/red scale showing commuting balance (blue = positive influx and red = outflux).

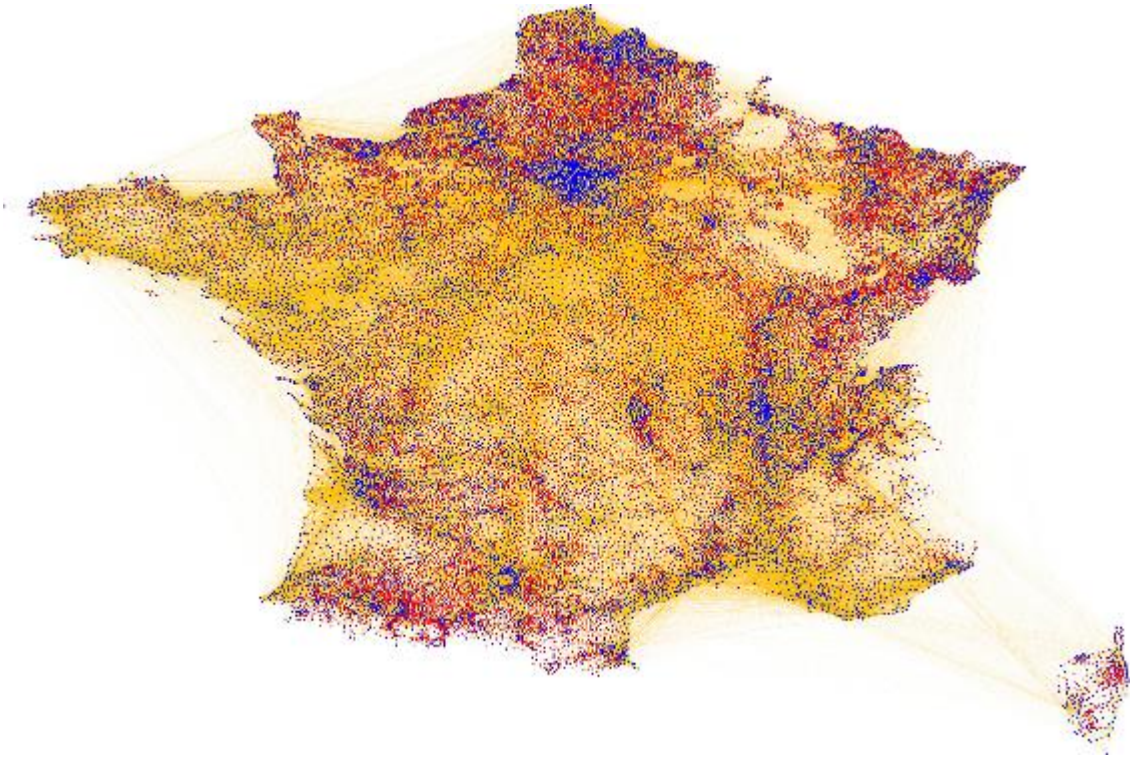


Figure 3 Influx-outflux of French commuters 2016

This model has been used as input to the COVID de-escalation policy advice report (De Groeve et al, 2020). The reports leverages synthetic population data to show some preliminary results at EU level on the relative contribution of sectors to COVID-19 transmission and economic impact. It identifies sectors with high epidemiological risk and high socio-economic importance (health, manufacturing, construction, accommodation and food services and wholesale/retail).

4.2 Activity-based modelling and behavioural data

We have used the synthetic French population to test several scenarios ranging from approximation of Eurobarometer data to house energy efficiency (to be published in a separate report) and the population data offered key advantages as the substrate for agent/activity-based modelling.

In practice, we have used a simplified approach using activity-based modelling, where the agents are just placeholders and we model their behaviour within a network of activities.

The model structure:

Agents – population in households in houses from the house-family mapping

Destinations – home, work, school, shops from work/school mobility mapping

Travels – based on HETUS[19] profiles.

The HETUS data for France contain 27.701 time diaries. The profiles are very detailed and contain place/activity data on 10-minute resolution, 1440 time slices per day. The data shows also information on presence of other family members which is very useful e.g. for distinguishing whether the reason for travel was to bring a child to school.

Another huge advantage of how the profiles were structured was information on day of the week, how normal this day was, illness, vacations etc. We could use other sources of data such as illness rate by economic sector to add another probability input to our model.

[19] Harmonised European Time Use Surveys (HETUS) https://ec.europa.eu/eurostat/web/time-use-surveys

There are also limitations in the HETUS data such as lack of information on regions, whether the data pertain to urban or rural population, from the Nord or South of France e tcetera. French statistical office INSEE is one of a few countries participating in HETUS that does not provide information on profession of the respondent. This led to a much weaker profile assignment since we only had economic sectors. In addition, no behavioural profiles were available for people under 17. There is no information on how typical or out-of-normal this behavioral profile for the respondent actually is.

Close inspection of all the profile data has shown many irregularities too: 1) Far too many people start too many activities precisely at XX:00 hours, i.e. they have filled the time-use diary in retrospection remembering only roughly the start of the activity (e.g. there are peaks on travel that starts at round hours). 2) Lifecycle (e.g. single person without children) in many cases does not fit the profiles (travel with children). In general, data do not seem to have passed rigorous tests.

However, these little nuances do not limit the usefulness of the HETUS data. The profiles are well balanced across lifecycle groups and provide wealth of behavioural information for analysis.

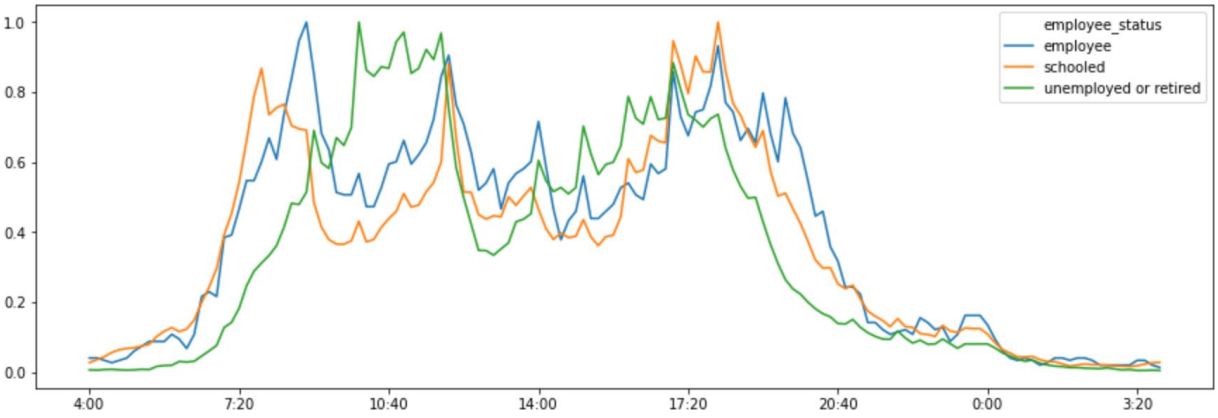


Figure 4 Example of aggregated time use diaries, here normalized for the viewer's convenience. "Activity" is defined as the fraction of people moving at a given hour.

We have mapped the HETUS behavioural profiles to our population. There was only a limited number of possible merging points that will still lead to A) full population coverage, B) enough entropy. Therefore we have selected:

- Gender,
- Age group (taken from HETUS, 10-17,18-19,20-24,25-29,30-34,35-39,40-44,45-49,50-54, 55-59,60-64,65-69,70-74,75-79, 80+'),
- lifecycle (HETUS compliant data were created from the INSEE data):
    - Below 25 years with no children < 18 years and living in parents' household,
    - 25 – 44 years with no children < 18 years and living in parents' household,
    - Below 45 in a couple (married/cohabiting) with no children < 18 years,
    - Below 45 with no children < 18 years and living in another household arrangement,
    - Single parent (all ages) youngest child <18 years,
    - (all ages) in couple (married/cohabiting) with youngest child 0 – 6 years,
    - (all ages) in couple (married/cohabiting) with, youngest child 7 - 17 years,
    - 45 - 64 in a couple (married/ cohabiting) with no children < 18 years,
    - 45 - 64 with no children < 18 years and living in another household arrangement (including those living in parents' households),
    - 65 and above in a couple (married/cohabiting) with no children < 18 years,
    - 65 and above with no children <18 years and living in another household arrangement
- NACE1D economic sector (1-8),
- Labour status (a mapping between the TACT variable from INSEE data with IND17_1 from HETUS).

We show in Figure 5 the distribution of the different lifecycles with respect to the age group:

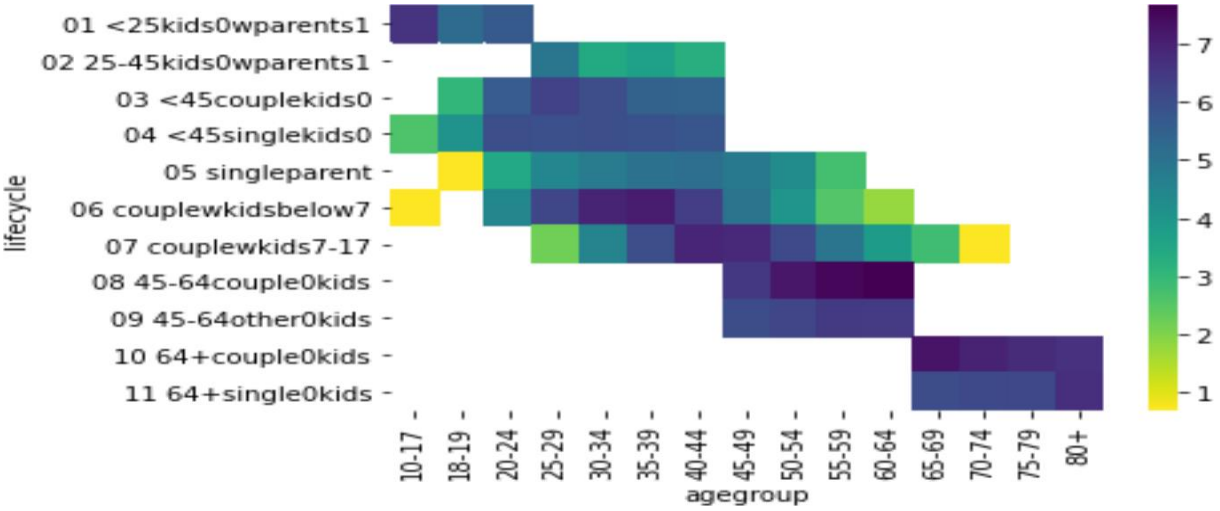


Figure 5 Distribution of lifecycle among age groups

The HETUS profiles are using statistical weights helping us to distinguish how important the profile is, and how high is the probability that this specific profile can be chosen for the specific person.

For simplification we have chosen to model a typical day of the population of Commune de Lille (59350)[20] consisting of 230.000 people plus people commuting into Lille. In total, we have identified 170.000 economically or schooled people that go to Lille to work or attend school. Furthermore, in order to utilize the highest variability of profiles, we have selected Tuesday and April as the most typical and data-rich time slices and added other profiles with adequately lower probabilities (e.g many people leave for their weekend houses and this can be seen in the Monday and Friday profiles) (Figure 6).

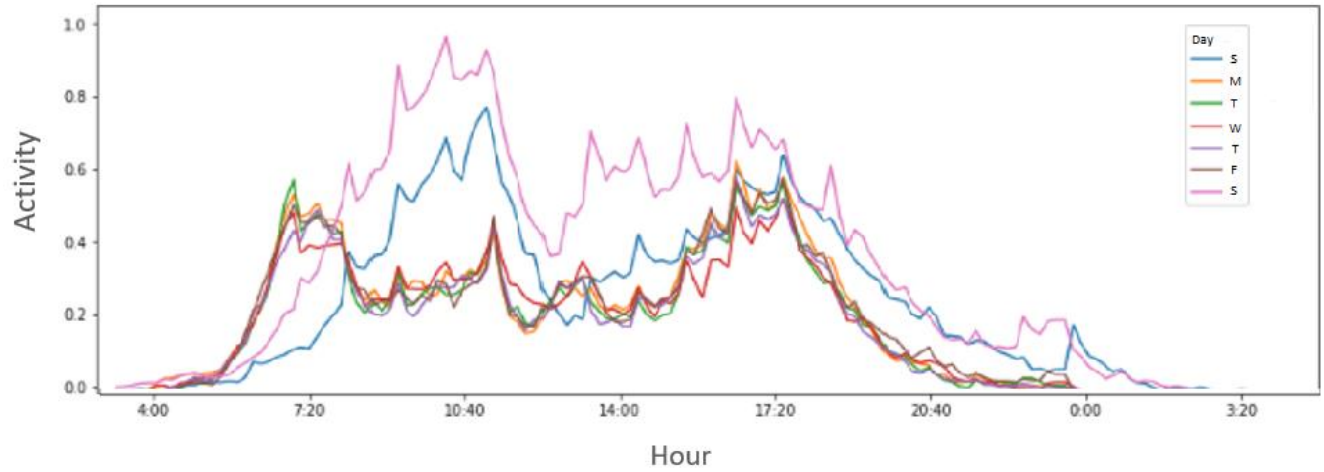


Figure 6 Aggregated time of travel by the day of the week. Please note that, in the legend, the week starts with Sunday.

In total, we have obtained 9000 categories for behavioural profiles, each consisting of at least 2 profiles. As the next step, we have started modelling the population. Below is the population distribution resulting from the assignment of households into real houses.

[20] https://www.insee.fr/fr/statistiques/2011101?geo=COM-59350

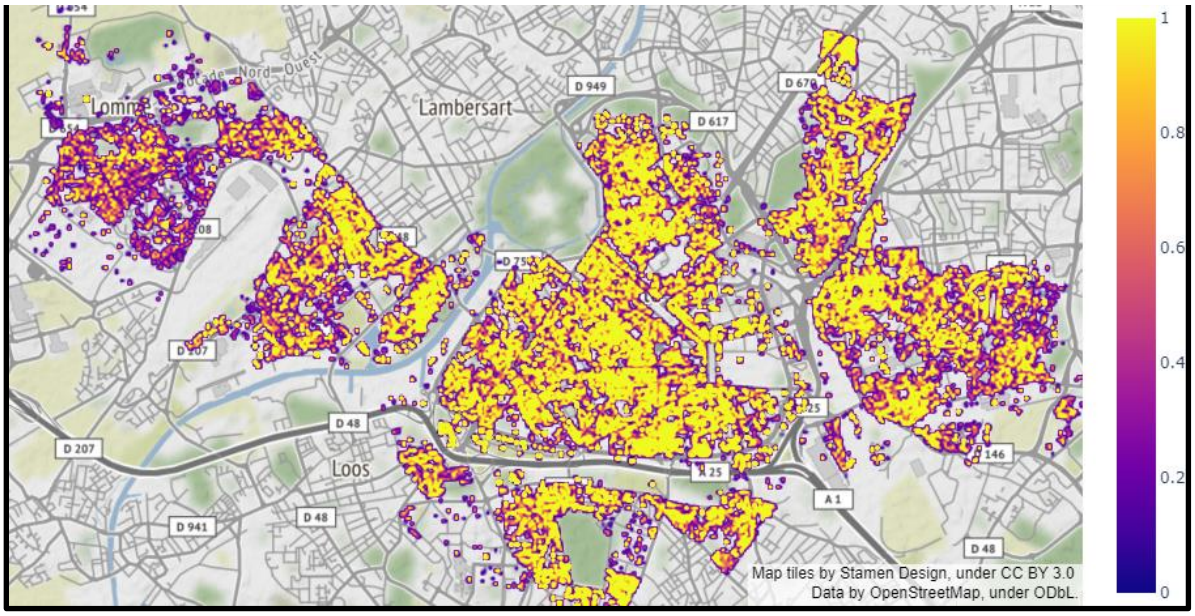


Figure 7 City of Lille simulated population distribution

The population distribution was quite fitting GHSL European Settlement Map R2019[21]. As the next step, we have added coordinates for work, school and shopping to the data. The assignment has been highly speculative, but gives a very good clue about people's destinations. In order to simplify calculations, for every travel a shortest path was calculated as if happening using a car or public transport.

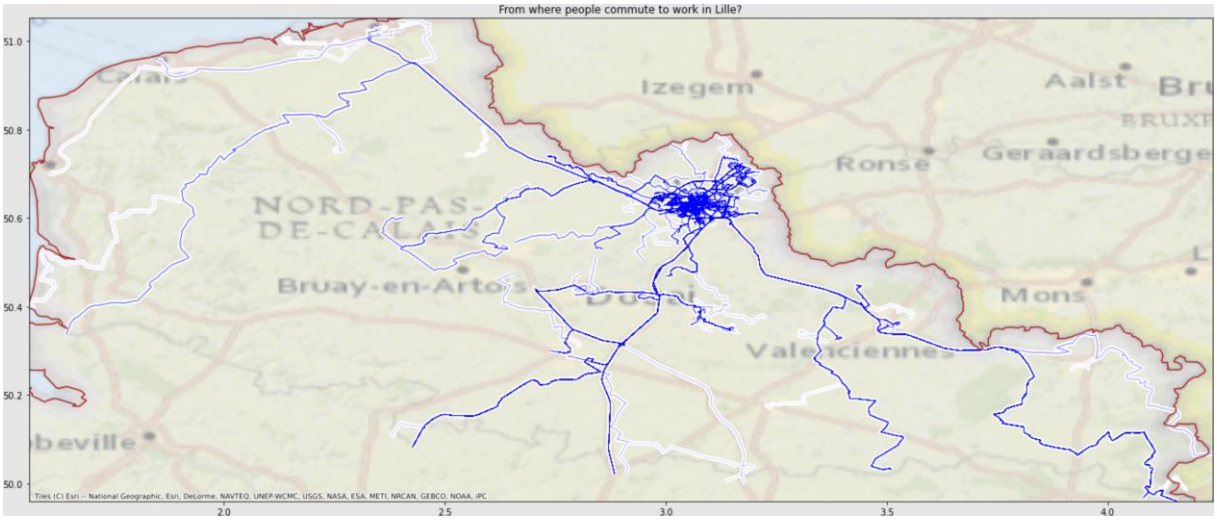


Figure 8 City of Lille regional inbound commuters pattern

The travel pattern assignment was run in pandas data frame by random weighted selection from equal profiles. To prepare a more robust simulation, the Monte Carlo method was used by repeating the assignment 10-times for every person, and all data generated have been used. For every 10-minute interval, a map was calculated, together with grid calculation showing whether the area cell has positive or negative inbound people flow. Total calculation time on a standard workstation was close to 3 hours. The randomized profile assignment takes about 10 minutes per run and visualisation about 1 hour. The routing was pre-calculated for all possible origin-destination pairs and took about 20 minutes. As we can see from the following images, the reason for travelling quite convincingly demonstrates the capabilities of this behavioural model. Before 7:00h, most of the travels are work related; before 8:00h, the school-related travel starts, which dominates also the time between 12:00h and 13:00h, as well as between 16:00h and 17:00h. After 19:00h, work mobility gets increasingly replaced by shopping/leisure travel patterns, which culminates by 22:00h. After 1:00h and 2:00h, all travel patterns diminish and the city goes to sleep.

[21] https://ghsl.jrc.ec.europa.eu/ESMVisualisation.php

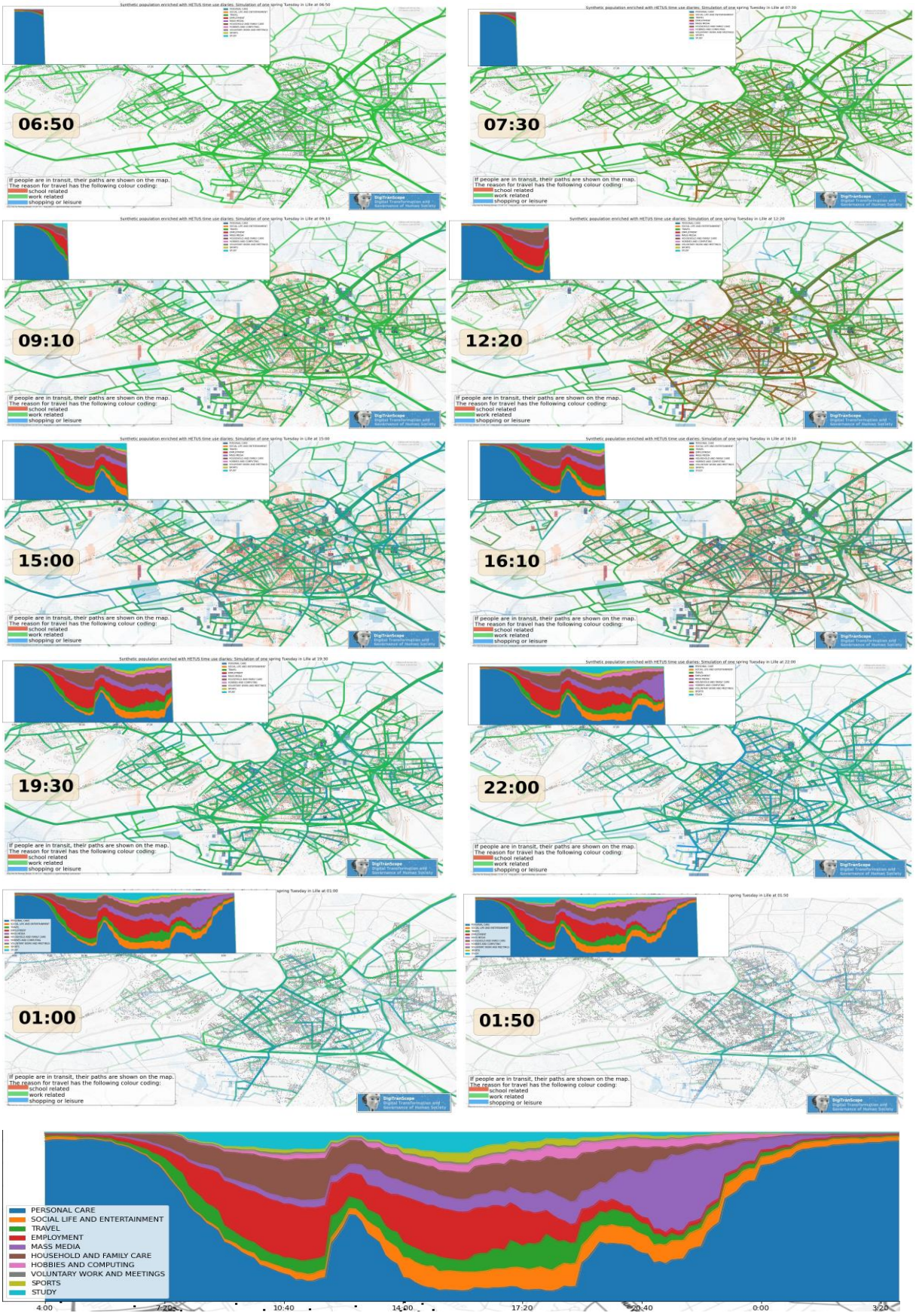


Figure 9 Activity-based model of city of Lille - Reasons for travel

Reasons for travel is only a trivial example of what can be achieved if we analyse the population as a graph together with mapped behaviour, be it work, consumption, sentiment or opinion mapping. We have no interest in knowing if and where the actual person went. But, when re-aggregated, the picture becomes rather consistent and informative.

The combination of realistic population with realistic behavioural model can become very useful for a range of computational models such as regional planning, pollution exposure assessment, or building energy models: as shown in Figure 10, mobility patterns can be combined with data on the household structure and on the building properties to construct an energy efficiency model encompassing all those inputs and giving a more realistic and granular depiction of energy use.

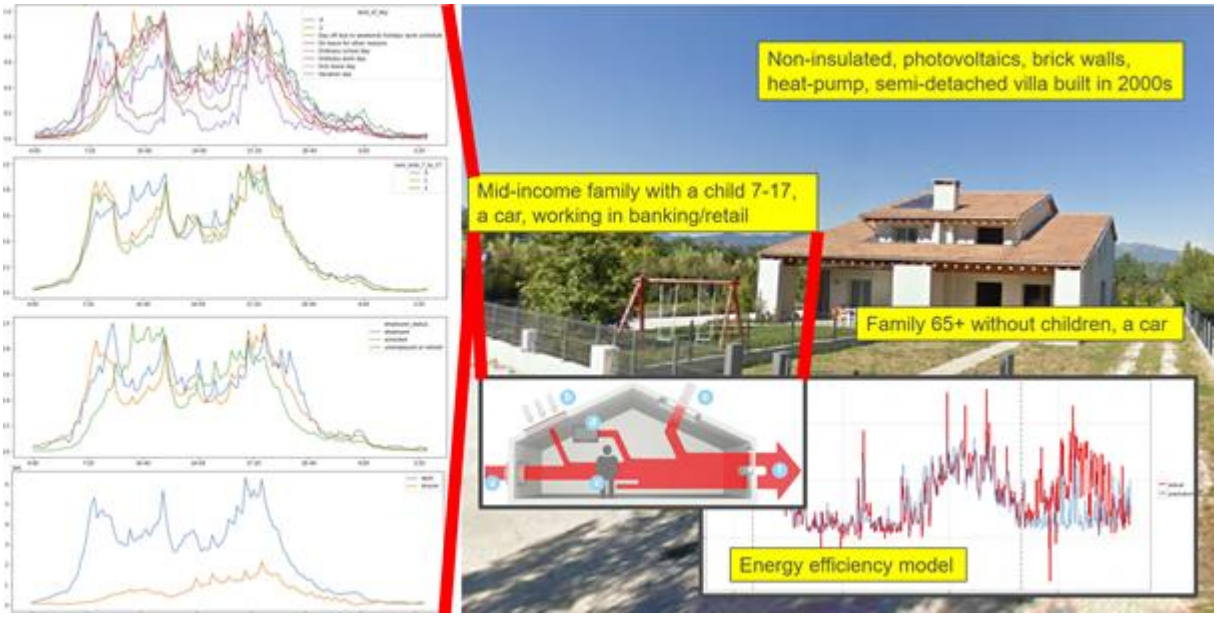


Figure 10 Example of inputs to building energy modelling

4.3 Lessons learnt

The synthetic population generated from the population model has been found very useful. The most important aspects are the quality and heterogeneity of population structure, and how easy it is to build agent- and activity-based models.

The method presented in this chapter for population generation is convenient due to its reuse of the upstream model provided by the French Statistics Office (INSEE). Population has been distributed into dwellings using limited available knowledge and more data would be needed to provide more precise assignment of families into houses, for instance the prices per square meter and the houses topology (surface, number of bedrooms) would help establishing the link between households in a specific income range and composition and houses.

Behavioural profiles from the HETUS data are a very useful guidance, though a closer inspection shows how unbalanced the profiles are. For research and demonstration purposes, the data are fully sufficient and open the path to quantitative behavioural modelling.

5 Use case 2: Comparing data aggregates, machine learning or synthetic population in the case of Amsterdam

In this use case, we have chosen Amsterdam as an example of comparison between three methods to analyse socio-economic and demographic data and make inferences to support policy making:

1. *statistics*
2. *robust analysis of statistical data using machine learning*
3. *Synthetic population.*

Our aim in this Section will be to investigate which inferences can be derived via the above methods and to which extent they are helpful in designing a policy targeting the energy efficiency of Amsterdam buildings.

The policy instrument under scrutiny belongs to the "Subsidies and arrangements for tenants and owners"[22] offered by the city of Amsterdam. These subsidies span from "Compensation for relocation and residential modifications" to "Subsidy for green roofs and green façades" to "Energy loan". We have picked the energy loan policy instrument to compare the aforementioned data analysis approaches, as this policy depends on population's capability to understand the need for implementation of the energy saving measures and their willingness to co-finance it. On the City of Amsterdam webpage, the instrument has been defined as:

> Owner-occupants, small-scale tenants and homeowners' associations in Amsterdam can take out a loan at very low interest rates to make energy-efficient improvements. This includes investments like double glazing and home insulation, but also solar panels, a heat pump, heat exchangers, or a greywater reuse system.

What is the energy loan from the municipality of Amsterdam?[23]

The energy loan is a loan with which you can make investments in energy-saving measures and sustainable energy for your house or houseboat. The loan has a fixed low interest rate.
Measures for which you can borrow:

- *applying insulation (floor, facade, cavity wall, roof, pipes)*
- *insulating HR ++ (+) glazing, if necessary including new windows / frames / doors*
- *purchase of solar panels, solar tiles, storage batteries or solar water heater*
- *LED lighting in general areas and outdoor lighting*
- *hotfill connection, shower heat recovery*
- *green piles (foundation piles) with soil exchangers*
- *heat pump*
- *gray water system*
- *green roof*

To be eligible for the energy loan, the measures must meet the conditions from the Sustainable Building List. Other values may apply to a monumental building. You can always discuss the options with the municipality.

You can apply for the loan if you fall into one of these groups:

- *owner-occupiers*
- *small private landlords (maximum eight homes in Amsterdam)*
- *Owners Association (HOA's, with ten homes or more)*
- *tenants*

These conditions apply to an energy loan:

- *The home must be an existing home or houseboat.*
- *The house must be inhabited permanently.*

We have explored the capabilities of the three different methods mentioned above (statistics, machine learning, synthetic population) to design effective policies contributing to the energy transition targets.

[22] https://www.amsterdam.nl/en/housing/subsidies-arrangements-tenants-owners/
[23] https://www.amsterdam.nl/veelgevraagd/?caseid=%7BED1B34DF-DF07-4D6F-81B4-D3E31F67B7FA%7D

We were not seeking to know how the actual policy was designed and approached the analysis without "a priori" hypothesis.

While Amsterdam has a 2016 population of 834,713[24] in the city limits, the urban area has a population estimated at 1.1 million and a greater metropolitan area with a population close to 1.6 million. We will focus on the inner city's 800k+ inhabitants. Over 40% live in social housing with uneven distribution, 70% properties are rented, there is ongoing gentrification with palpable increase in house prices.

5.1 Statistics

For the standard policy design, we will focus on the available facts from the City of Amsterdam Open Data Portal[25]. In Amsterdam, there are 101684 buildings, out of which 96984 have status "in use".

We have selected the building of these types:

terraced house	34189
apartment middle	33693
apartment low	14562
corner house	9077
detached house	2436
semi-detached house	2299
apartment high	728

Table 3 Building types in Amsterdam

Houseboats and caravans never have the status "in use" in the data provided.

From Table 4 we can see how the heat consumption rapidly increases with the years except for combined heat and power, hence immediate action seems to be justified. Identification of household actors who have to implement the policy: 251411 single person households, 100356 couples with no kids, 74991 couples with kids, 40884 single parents, 5876 others. The average yearly net income of a single person household in Amsterdam is 23k EUR, a couple with no kids 46kEUR, a couple with kids 65k EUR, single parent household 35k EUR.

Private and commercial heat consumption (TJ), 2014-2018						
	2014	2015	2016	2017	2018	
heat supply cogeneration plant Diemen						
private consumption	339	381	403	431	505	
business consumption	757	806	964	974	912	
total	1095	1188	1367	1405	1417	
heat supply Waste Energy Company						
private consumption	129	156	172	251	296	
business consumption	295	336	374	418	520	
total	424	491	546	669	817	
combined heat and power						
private consumption	230	226	285	182	207	
business consumption	82	83	80	42	44	
total	311	309	365	225	251	
total heat consumption	1831	1988	2278	2299	2485	
source: Department of Space and Sustainability (CE Delft)						

Table 4 Private and commercial heat consumption in NL

From this perspective, a loan in the range of 2500-15000 EUR with negligible 1.06% interest rate (below standard interest ratio as motivation) seems perfectly affordable.

[24] https://data.amsterdam.nl/specials/dataverhaal/de-amsterdamse-bevolking-sinds-1900/c7e4ae3c-0808-4d7d-a3a1-1c4a72a53377/

[25] https://data.amsterdam.nl/

5.2 Machine learning

We can imagine that if the analysis we made above was presented to policy makers, it would be convincing to them. There are some undeniable facts and trends, the instrument is affordable to mainstream actors, and everyone has been fully informed about the opportunity.

But what happens when everything is set, money is there, but very few people utilize the loan instrument? What could be a barrier in this case?

For instance, we can hypothesize that income could be such a barrier: 42% of households have low income, 15% on or below social minimum for which such loan could still be difficult to afford.

Therefore, for further scrutiny on this interplay between energy consumption and socio-economic factors, we have used two datasets. “Basisregistraties Adressen en Gebouwen” (BAG), contains information about addresses and buildings in the Netherlands and is freely available at the Land Registry (Kadaster, n.d.). Next, “PICO” is an energy information portal that provides both open and private data on energy consumption and building characteristics. The building characteristics are derived from the height map “Algemeen Hoogtebestand Nederland” (AHN) (Geodan, n.d.), and the solar generating potential is calculated from the datasets BAG and AHN. However, neither dataset describes which family occupies which house, what is the household structure and income, or who are the household members.

1.1.1 PICO energy portal dataset

To understand whether there is potential to uncover the correlation between socio-economic factors and gas consumption in the data, we have used 2-dimensional UMAP projection of the complete PICO data on the 17,000 buurt areas[26] with data on gas consumption, inhabitants age and income percentage, house construction year and type, and ownership type. As we deal with high dimensional data points, UMAP projection allows us to, nevertheless, visualize clusters in a low-dimensional (2D in this case) representation,

High-consumption areas (>1500) have been highlighted in red color, low consumption (<300) in green color (Figure 11).

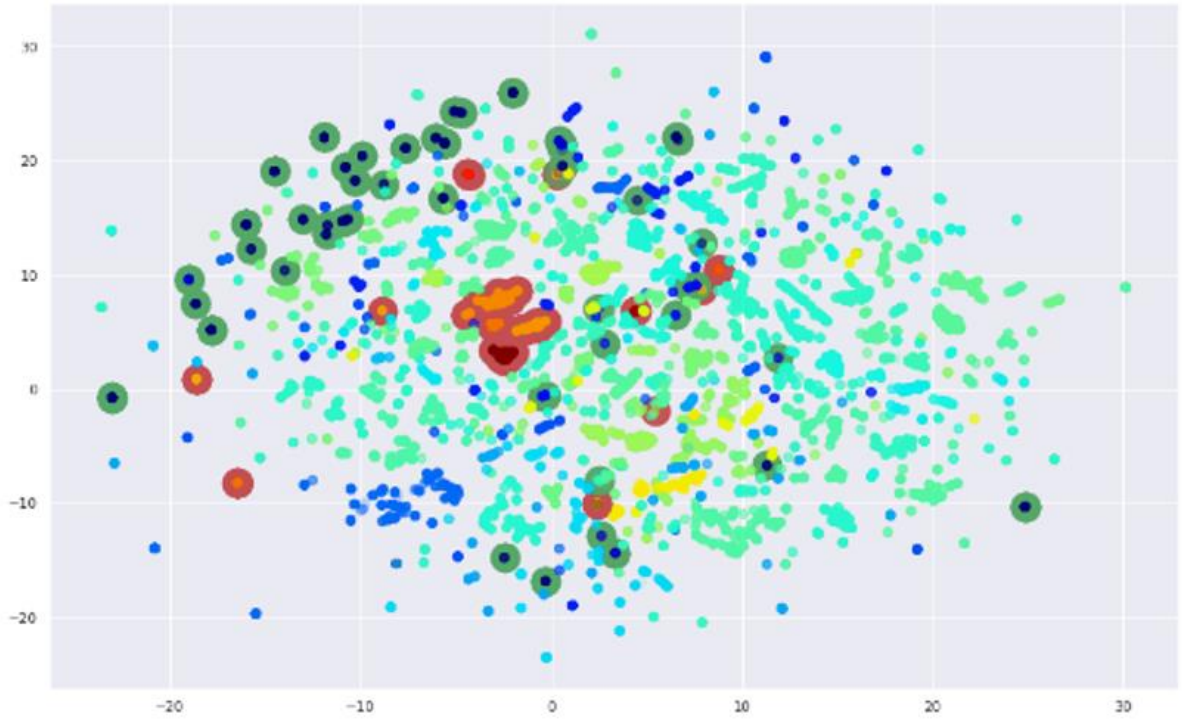


Figure 11 Household gas consumption-related energy efficiency multivariate projection

[26] Neighborhoods characterized by a 6-digit postal code. CBS Kerncijfers wijken en buurten 2016, https://opendata.cbs.nl/statline/#/CBS/nl/dataset/83487NED/table?dl=3E45

We can see in Figure 11 that low and high gas consumption areas clustered, which encourages further exploration to find variables with major impact.

As displayed in Figure 12, data for each housing type show three distinct patterns when drawn using two-dimensional kernel density estimation of distribution of average area gas consumption vs. consumption of households in the area by house type. Data have been reweighted to reflect the real distribution.

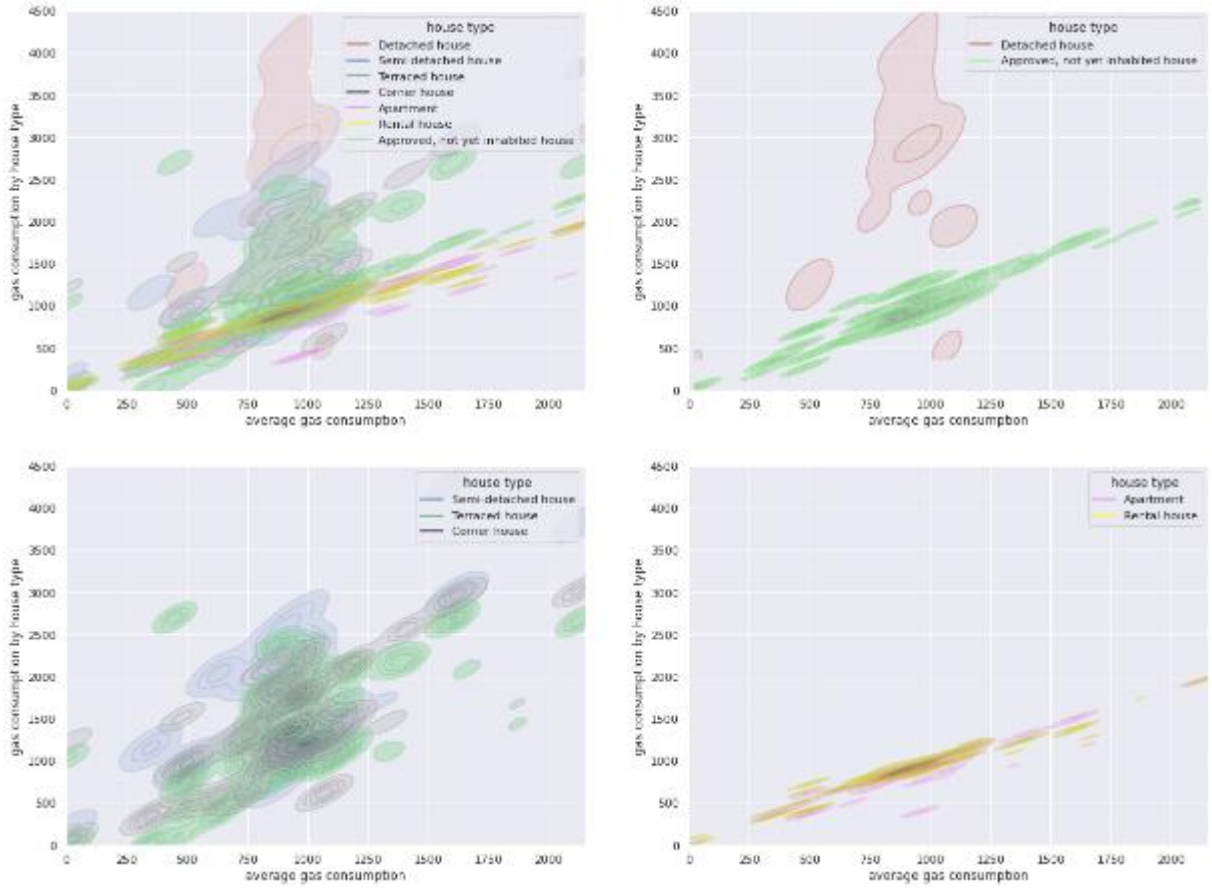


Figure 12 Average gas consumption by house type gas consumption. From top left corner, clockwise: a) kernel density estimation of all house types, b) only Detached and Approved not yet inhabited houses, c) only Semi-detached, Terraced and Corner houses, d) only Apartment and Rental houses.

The three major patterns are clearly visible here. First, detached houses are usually major contributors to the area consumption (Figure 12 b), except for low energy or probably passive ones at the bottom of the chart. Apartment, rental houses as well as approved but not yet inhabited houses tend to have very low gas consumption (Figure 12 b-d). Row houses are coming in full gamut, from the very energy efficient ones to the other side of the spectrum.

To further investigate such quantitative links, we studied a simple Pearson correlation in 17060 buurt areas of Amsterdam with non-zero gas consumption, the rationale behind being that these correlations would pinpoint the socio-economic variables that contribute the most to gas consumption.

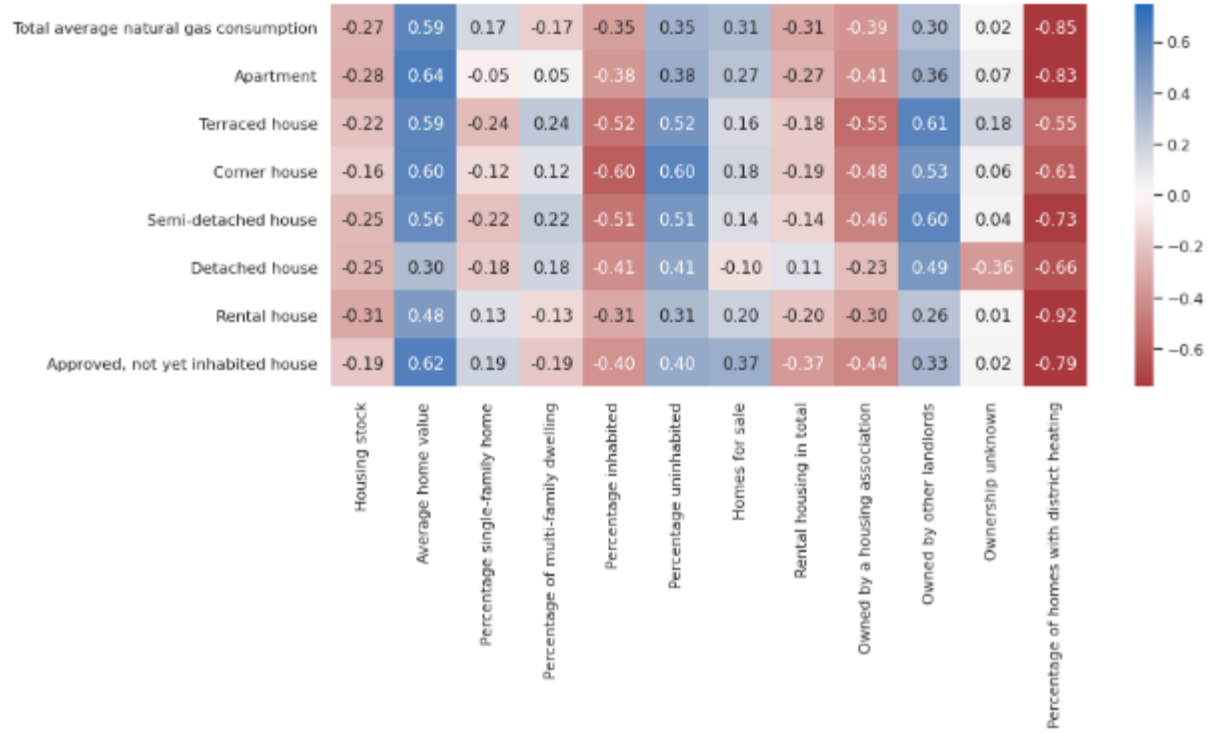


Figure 13 Correlations between house type and gas consumption

Indeed, as we can see in Figure 13, more costly houses tend to have larger surface area and it justifies the correlation with gas consumption. Interestingly, detached houses have smaller energy footprint, most probably due to massive private investments into energy efficiency. A percentage of inhabited houses correlates with detached houses as they have relatively higher consumption but has negative correlation with apartments and terraced housing, where the higher number of occupied apartments leads to lower energy wasted through the outside walls and walls with uninhabited apartments. The presence of homes for sale is surprisingly highly contributing to the average energy consumption. We can see how large apartment houses owned by associations and rental housing (e.g. hotels) have already been insulated. We can say we should focus on privately owned houses, either free standing or in a row. Unsurprisingly, district heating which allows using the gas just for cooking leads to a huge slump in the reported gas use.

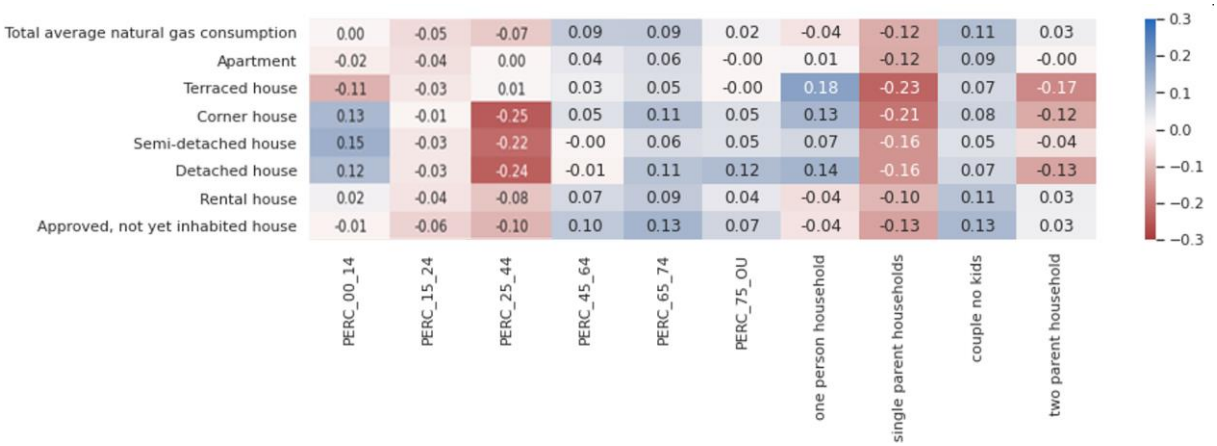


Figure 14 Correlations between demographic composition and the gas consumption by housing type. On the x axis, the variable "Perc_MIN_MAX" indicates the demographic group with an age in the interval [MIN, MAX).

The preview findings have been further reinforced in Figure 14. Data show that people of age between 15 and 44 years tend to live in more energy efficient houses. The older people the less energy efficient housing. Apartment buildings are more efficient yet elder generation still uses more energy. Children in

household increase the gas consumption, yet the single parent families cannot afford the same thermal luxury.

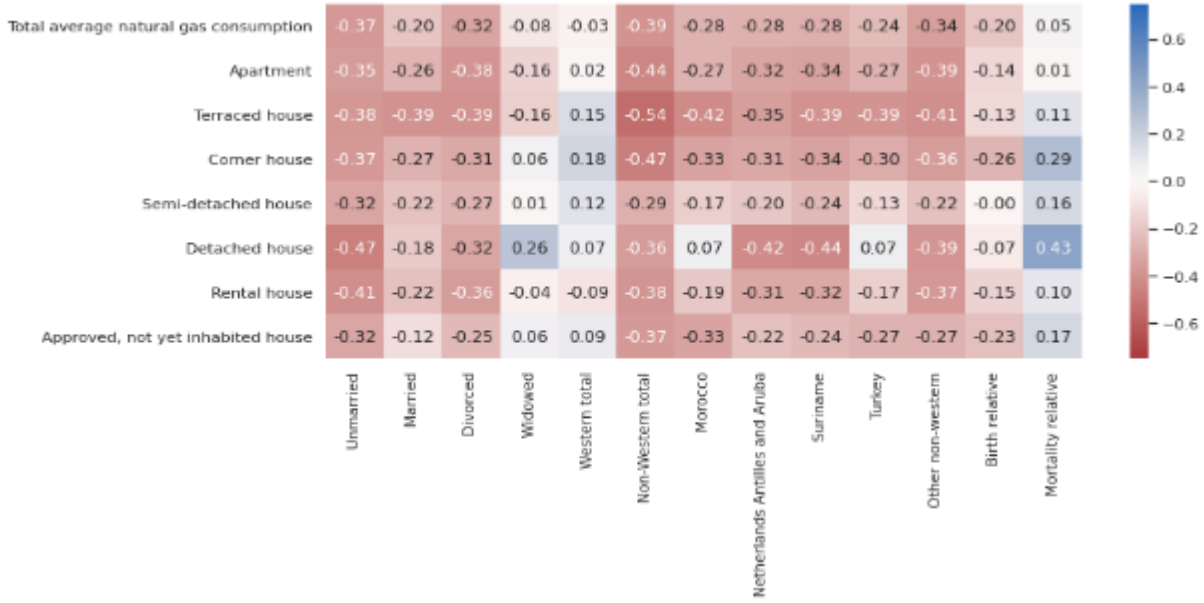


Figure 15 House type gas consumption by ethnicity and morbidity

Exploring the details of mortality and ethnicity in Figure 15, we can see that people of Western origin tend to consume substantially more than people of other ethnic origins. Moroccans and Turks are possibly the better situated among other ethnic groups and once they reach the wealth allowing them to afford their own houses, they do consume energy at the level similar to Westerners. Alternation of generations seems to also have a big impact on the gas consumption.

Let us not forget here that correlation does not induce causation. We may point to group behaviour but we have no information about individual behaviour. Ethnical correlation may have origin in a type of building, specific underlying issues etc. Therefore, we would like to raise the caveat that we are just using one type of data without cross check, only the first input for a policy analyst.

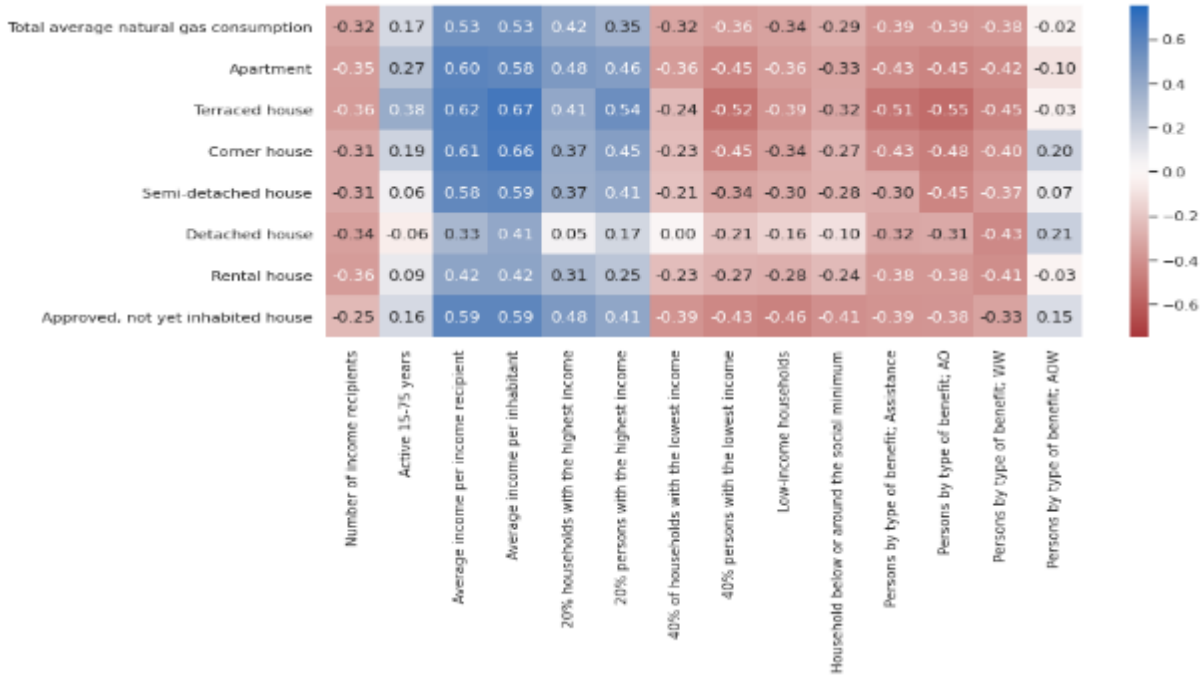


Figure 16 Gas consumption by income group

Going one step further, to the correlation with groups of different incomes, in Figure 16 we see a rather natural correlation between higher incomes and higher consumption. The more a person is dependent on state assistance, the lower the energy consumption. But here comes the paradox – are they living in better insulated houses or are they heating less because they cannot afford it? Obviously the latter, because the more people with the lowest income level (e.g. two unemployed people in the household) the lower the energy consumption. We need to normalize the data against the income level to truly understand the need for a policy instrument.

We have calculated the importance of features contributing to increased gas consumption over the whole dataset with the results shown in Figure 17, where we see that variables linked to income are the most influential with respect to the gas consumption.

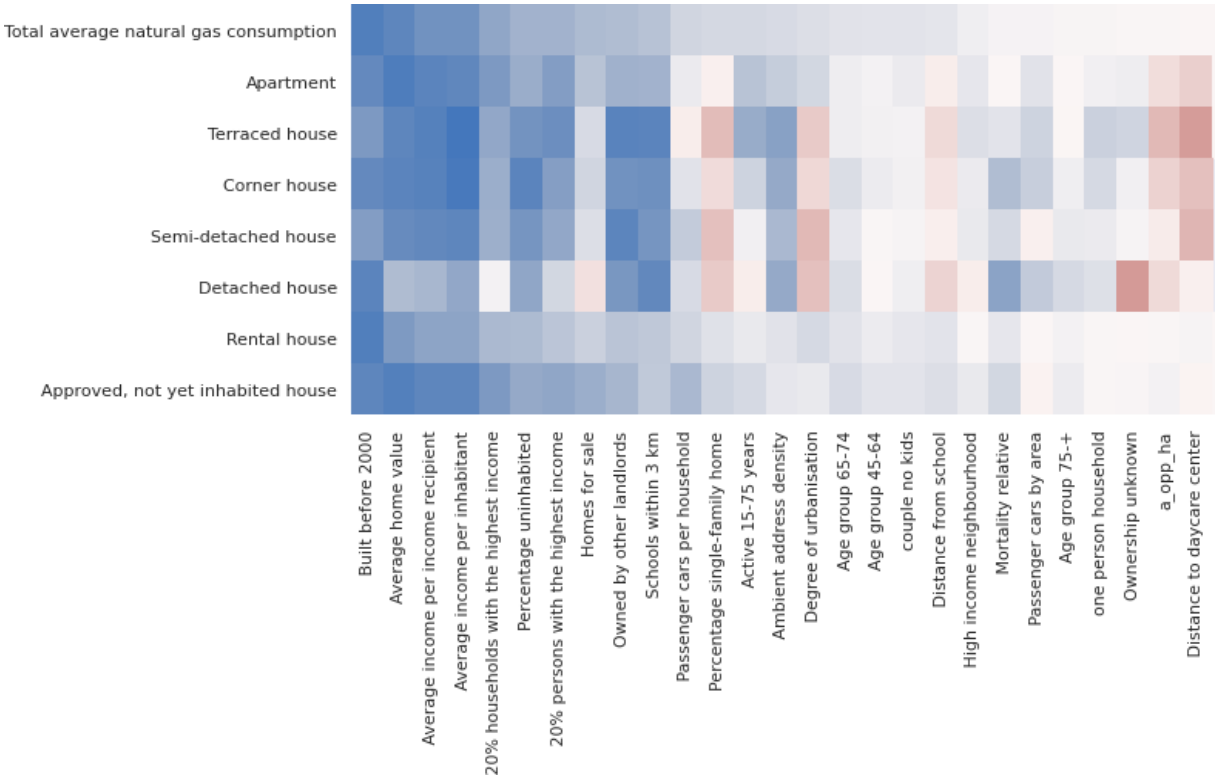


Figure 17 All key net positive correlations

At this moment of analysis it seems that the policy instrument has been designed to solve absolute numbers and not the real issue – households at heating poverty actually are not more energy efficient but since these families heat less they appear to be more efficient. This instrument thus widens the heating gap instead of solving it. Another policy, called "From big to better" funding scheme[27], incentivising moving from large but heat inefficient houses to smaller but better insulated ones, partially helps but only to a certain extent.

The age group of 75+ (and partially 45-64+) seems very much demonstrating this problem. As data show, the economically inactive people have much lower energy footprint; yet, the age group 75+ presence in the buurt area increases the average energy consumption. This leads us to an obvious conclusion that the age group 75+ lives in highly energy inefficient houses. Taking the loan to be repaid in a decade seems very little motivating leaving these people at much worse position. Unfortunately, even the data show that a higher death rate improves the energy efficiency of the area.

From Figure 17, we hypothesize that the income level has a higher impact on energy consumption than the type of house, the year when constructed or the age of residents and, therefore, the need for house insulation cannot be determined through energy consumption while energy poverty actually could play a major role. However, since correlation does not imply causation, we had to better test the data on a different dataset.

[27] https://www.woningnetregioamsterdam.nl/Paginas/Van%20Groot%20naar%20Beter

1.1.2 Household level energy estimation – the BAG dataset

Data in the BAG dataset contain information on individual buildings, their age, the residential function, etc. We can find a more precise link between the residential occupied building and the buurt area (6 digit postal code) gas consumption. The gas consumption is calculated differently in this dataset, it starts from zero and goes up to 17,000 m3/year. We had to approach the data differently.

To understand our data, the best plot is on a map where we see the data we have available. The white-to-red patches are the buurt areas with average-per-household gas consumption. We have buildings with a building type and we have one or more households per building.

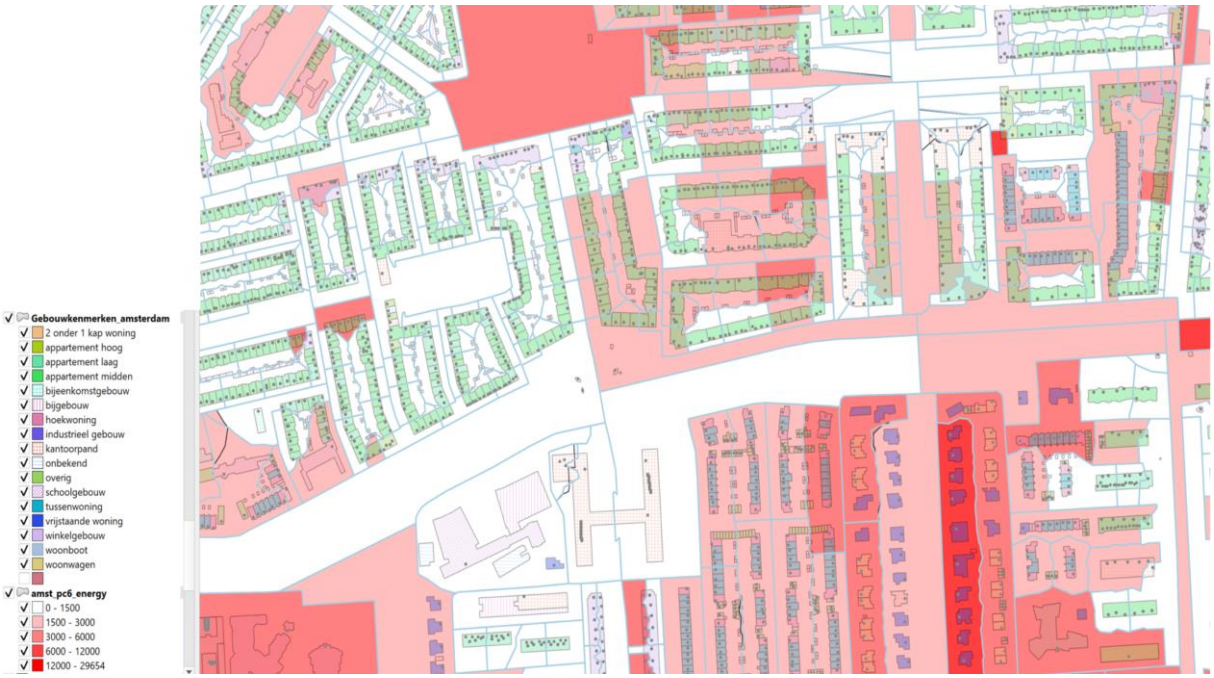


Figure 18 Average gas consumption per building type.

Figure 19 Average gas consumption per building type.

A careful revision of the data shows the low gas consumption buurt areas are either those with massive development, or where there is a high percentage of homes with district heating, or when no longer use gas for heating. Typically, these areas are in the south-west of Amsterdam:

Figure 20 Map of Amsterdam with yearly use of gas. White patches are areas with average gas consumption lower than 200m3/year

Since we do not have the actual gas consumption by household, building type and construction age, we needed to resort to find relations through proxy data. We have split the construction years into categories (pre 1945, 1946-64, 1965-74, 1975-1991, 1992-2009, 2010+) based on the construction law requiring different levels of house insulation. Consequently, we have selected variables using univariate SelectKBest algorithm with X^2 scoring:

	Specs	Score	# gas consumption: high, low
30	HOOGINKOME	133679	# Families with high income
29	LAAGINKOME	88963	# Families with low income
19	NIETWESTER	72374	# Percentage of non-Western origin people
31	UITKERINGS	68887	# Percentage of people with social benefits
32	ZELFSTANDI	60918	# Independently standing house
2	1965-1974	2752	# Percentage of houses built in 1965-1974
5	2010+	2217	# Percentage of houses built after 2010
4	1992-2009	2115	# Percentage of houses built in 1992-2009
1	1946-1964	2105	# Percentage of houses built in 1946-1964
3	1975-1991	1678	# Percentage of houses built in 1975-1991

Table 5 Scores of different variables with respect to the gas consumption.

We can see that in this dataset there is much higher importance of income level and percentage of non-Western origin inhabitants. The dependence on social benefits has practically the same importance as whether the house is standing independently. Only then come variables for house construction age.

We have studied the distributions in order to understand the relations between inhabitant gas consumption dependent on age structure and house types (Figure 21-22). The mapping on small-scale areas (buurt) by prevalent housing type (>40%) demonstrates very low energy efficiency of houses from the 1965-1974 and 1975-1991 periods.

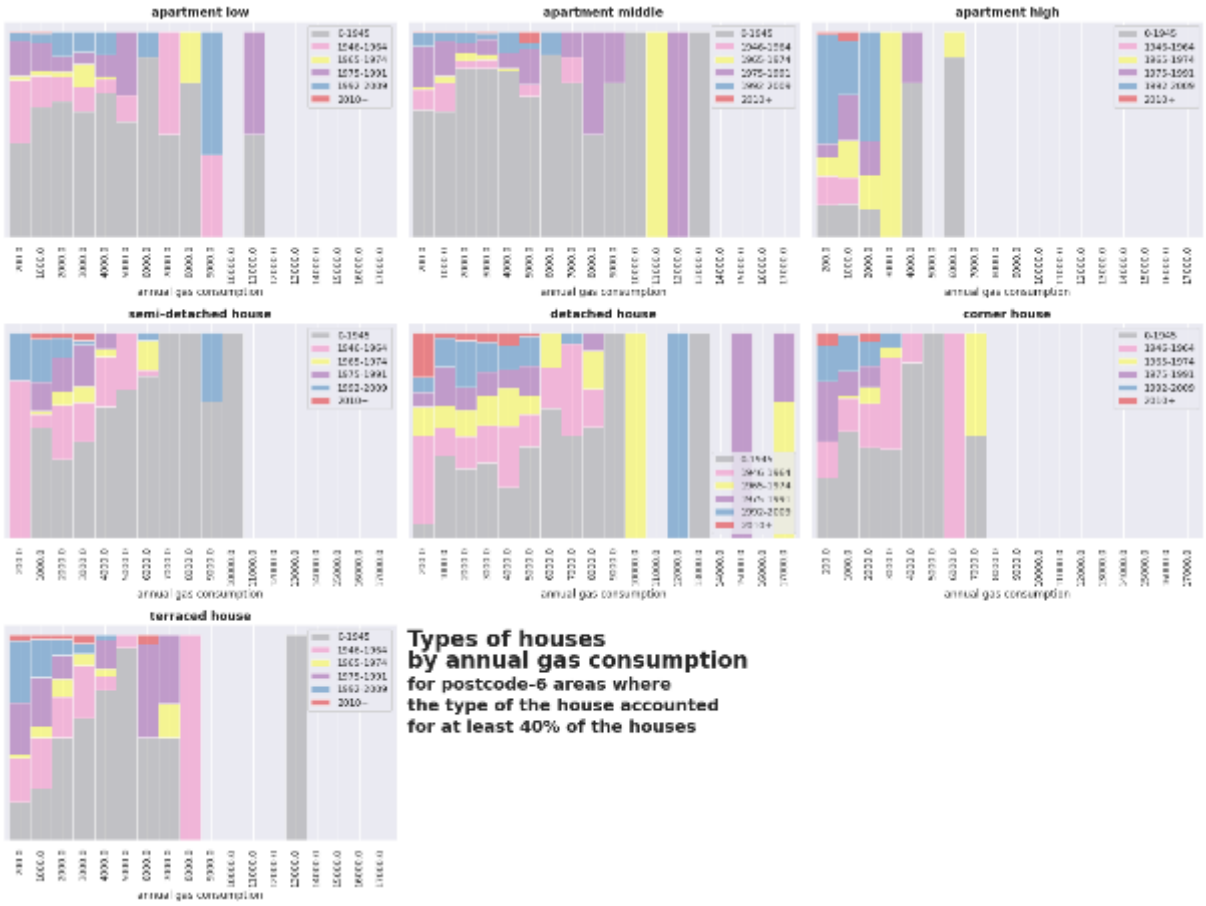


Figure 21 Relation between house type, year of construction and gas consumption

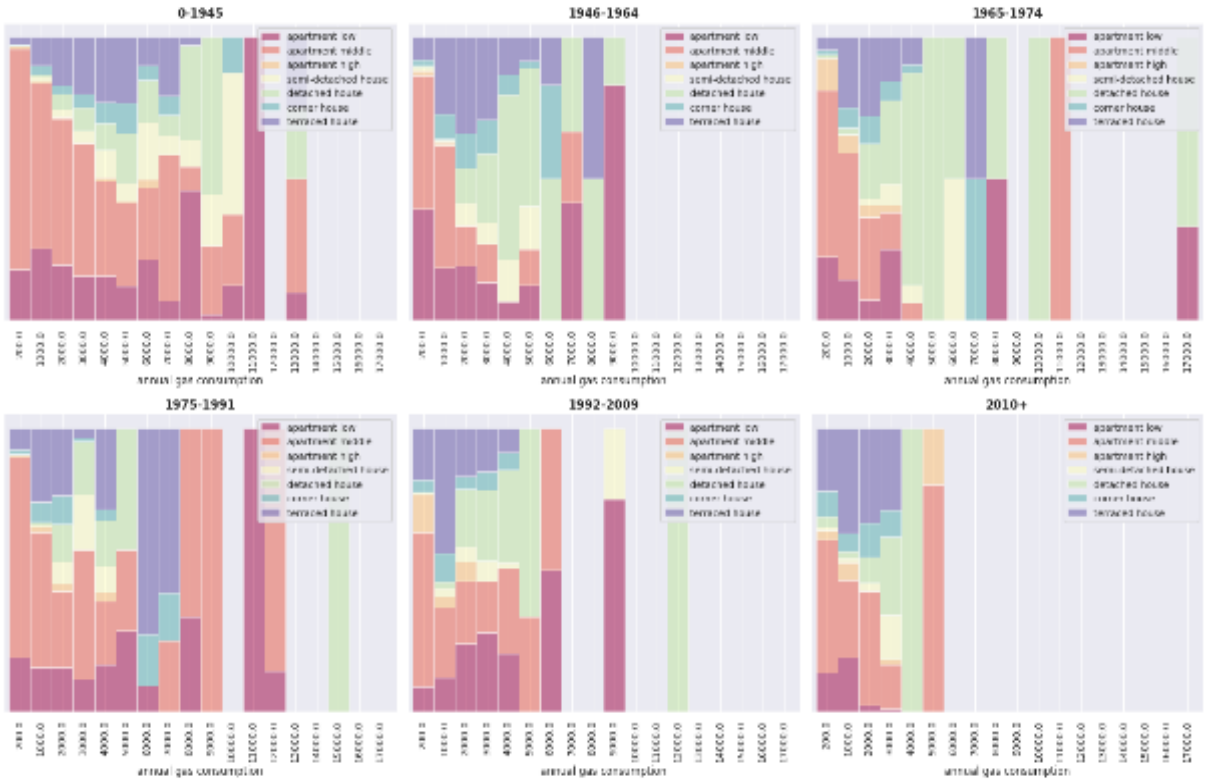


Figure 22 Relation between year of construction, house type and gas consumption

The assessment of the population distribution is substantially more complicated. We have utilized kernel density estimation to better understand distributions across areas (Figure 23).

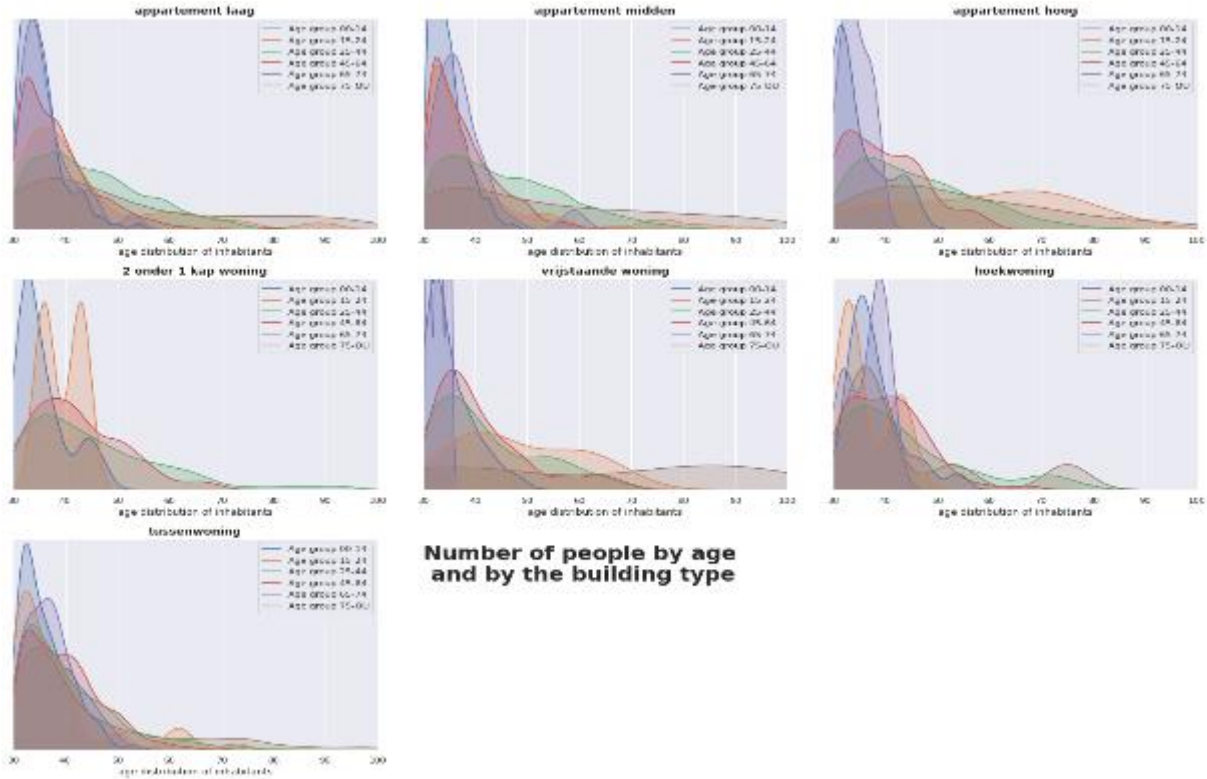


Figure 23 Kernel density estimation of households by age and building type

As we can see from the data, the elder generation often lives in the independent houses built in 1960s and 1970s where the energy efficiency suffers.

At this point, we have a good understanding of possible impacts and can build a complete regression model that models gas consumption through the variables linked to economic status, demographics and building characteristics.

The distribution of the average gas consumption in a 6-digits postcode area in 2017 (GASKV2017) is very slim and skewed and machine learning models would have troubles to build reasonably performing estimators. We have applied base 20 logarithm to rectify the situation (Figure 24).

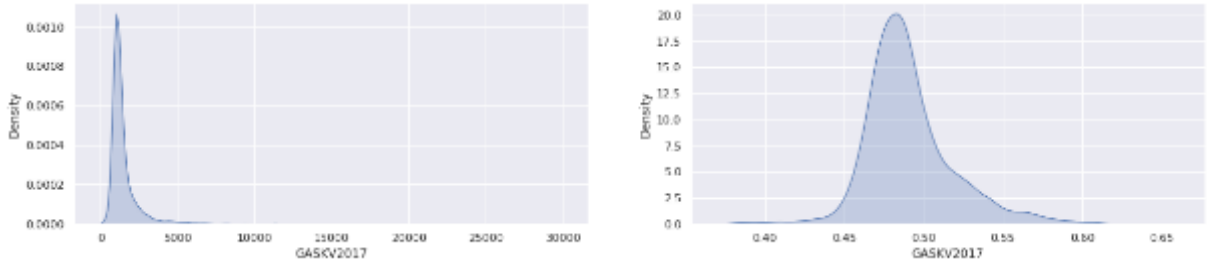


Figure 24 Original and logarithmic distribution of gas consumption

We ran a series of representative estimators – LinearRegressor, XGBoost, Random Forest, PLS, Ridge, KNeighborsRegressor. The high degree of error from all models outside 500-2000 m^3/year zone pointed to lack of data, possible methodological issues in data collection, and that we lack features that will truly shed light on the cause and effect.

The far best performing regressor was XGBoost and we have used GridSearchCV for exhausting hyperparameter search. The resulting estimator was rather satisfactorily:

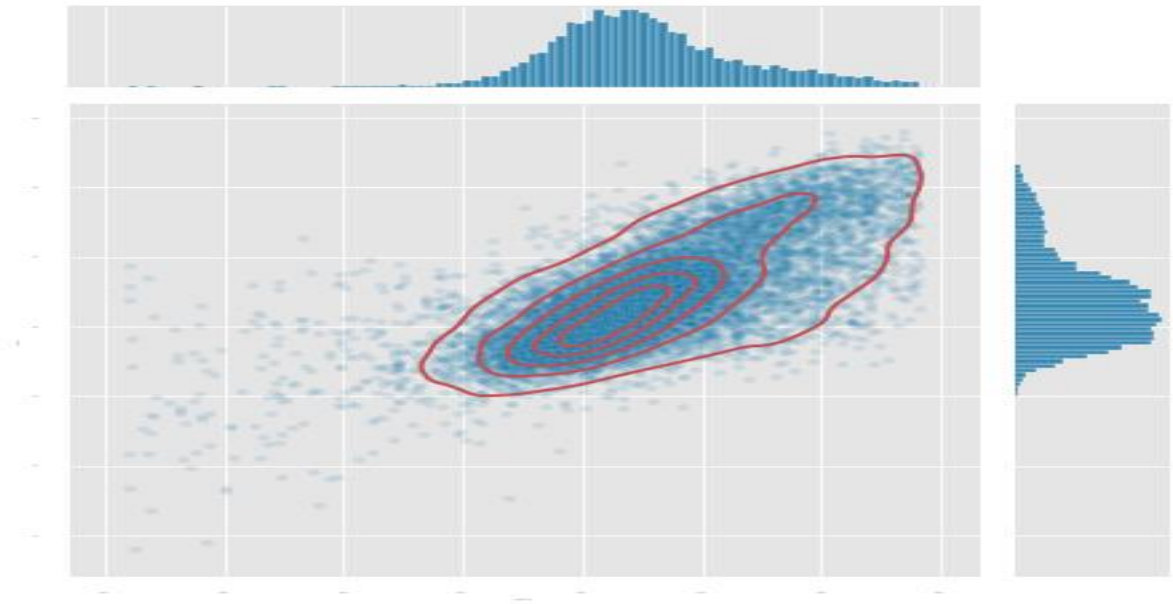


Figure 25 Ground truth value vs predicted value – the average gas consumption

The estimator worked best for values in 700-3000 m^3/year range, for values lower than 700 m^3/year we see a higher spread. This is most probably caused by higher number of houses not dependent on gas heating and using the gas only for cooking.

The sensitivity analysis shows variables by their identified importance in decision trees:

	25	33	7	26	19	21	15	1
name	Families with high income	Taxable salary	Families with low income	Independently standing house	Percentage of non-Western origin people	One Person household	PERC_25_44	1946-1964
val	0.0725079	0.0676837	0.0561656	0.0515327	0.0407464	0.0355845	0.0351836	0.0348434

Table 6 Sensitivity analysis if of the socio-economic variables in the decision tree model.

There is only some 2,400 independent houses out of 101,000 houses in Amsterdam, yet they have a huge impact on the resulting average gas consumption in the buurt area.

The income level had substantially a higher impact on energy consumption than the type of house, the year when constructed or the age of residents. This further adds to the suspicion that the need for house insulation cannot be determined through energy consumption and that this is the energy poverty that actually plays the major role.

5.3 Synthetic population

We have learned the possible influences of higher gas consumption. In the next step, we will demonstrate how synthetic population can be used to identify communities of common behaviour and needs. We will try to pick up the groups of inhabitants in such a way that the policy analyst can understand sociodemographic barriers to policy implementation.

So far, we have found that the high-income population lives in independent high-consumption houses that account for 2.5% of the housing stock. We have a suspicion that low-income population may live in non-insulated houses but energy poverty does not allow direct identification of the issue. We will try to validate or invalidate these generic signals in the disaggregated population.

We have reconstructed the population by applying multivariate distributions learnt from the city-wide data using machine learning. For example the university-level education of the Amsterdam inhabitants (distribution basic 38%, middle 29%, university degree 33%) is heavily leaning towards younger population (50% of university educated people are under 35 years).

The key difference between data aggregates and the individual data lies in the loss of structural information in the aggregated data. Disaggregated synthetic population serves us as a vehicle in stochastic mapping of a very complex function. By assigning these generated individuals to synthetic households and then to real houses, we have enabled the follow up analyses to approximate the function of gas consumption without having to explicitly express the function.

We have disaggregated population statistics using iterative proportional fitting (IPF) with applied above described multivariate distributions. The resulting population was created with high-entropy target, i.e. avoid grouping data around the median. Higher jitter could lead to higher error in correct population assessment but highly improve results of the analysis since we are not interested in individuals but in their communities.

The source data cluster well so there is a good chance that we will be able to find links in the data:

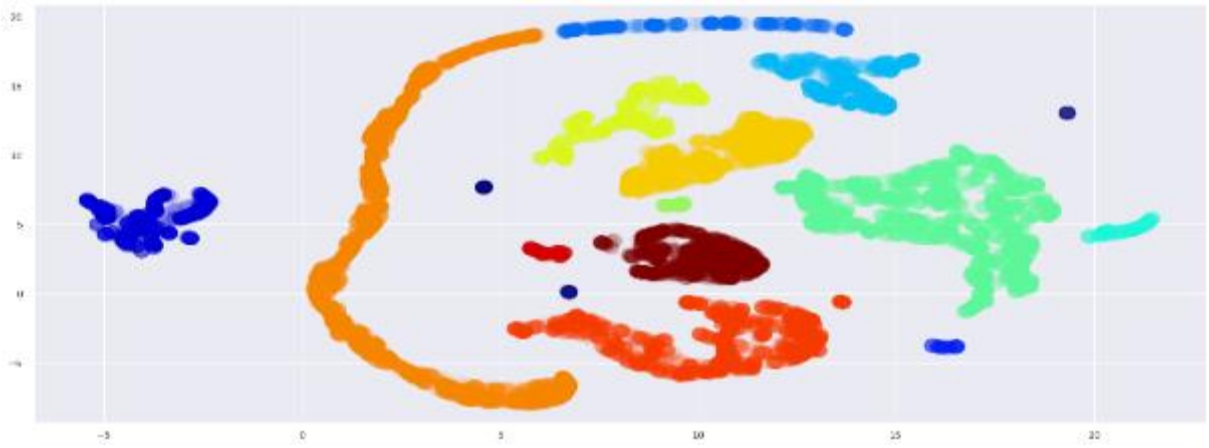


Figure 26 Source data clusters

Our analysis so far hints that low-income families with children may lack the means to invest into insulation, and, also, elderly people may have little interest in taking a loan, which spans several year, to insulate their houses. This evidence from the data could be the basis for a targeted survey to confirm these behaviours and preferences.

Our goal was to validate these findings from aggregated level against the synthetic individual data. The data of individuals, households and houses were joined and pivoted against gas consumption normalized by the apartment surface area. Even though the synthetic population was generated from the official univariate statistics and the distributions were learnt from the overall picture of the whole city, we can see the more nuanced yet credible outcomes of simple tri-variate comparisons.

For the validation study we have used 377,307 synthetic households of 597.608 people, 105.193 houses, 223.352 apartments in 16,323 census areas where all population, housing and energy consumption data were known.

Building type	gas_per_m2	count	year of construction
apartment middle	9.014925	310479	1917.919673
apartment low bldng	9.384372	97617	1899.477110
terraced house	24.763711	62859	1953.493820
apartment highriser	3.185430	29990	1963.295765
corner house	25.095076	16520	1964.857022
office buildings	4.349933	13957	1908.938884
industrial building	5.759316	12772	1872.515581
Outbuilding	24.664477	4574	1805.325754
detached house	25.232928	4159	1949.138735
semi-detached house	30.356774	4122	1951.295488

Table 7 Distribution of households in the study area and their average gas consumption per m^2 and year

To see if there are any distinct patterns of gas consumption noticeable in people's attributes, we have cut the synthetic individuals' records into 10 quantiles by their gas consumption. Binning is arbitrary high to see the resulting clusters for further guidance.

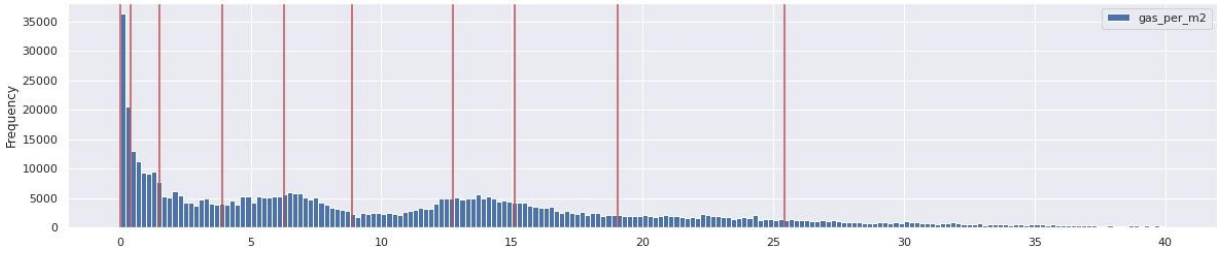


Figure 27 Quantiles of normalized gas consumption

A simple frequency analysis of these five groups yielded a simplified table with values of highest cardinality attributes for each target group (the variable "bin").

bin	PC3_code	construction_year	roof_type	education	age	social_assistence_dependent	income_household_interval	household_type	Building type
(0.00209, 0.405]	106	1965-1974	Flat	edu_basic	25_44	False	(23800.0, 43850.0]	two_parent_family	apartment middle
(0.405, 1.518]	106	1992-2009	Mixed	edu_basic	25_44	False	(23800.0, 43850.0]	two_parent_family	apartment middle
(1.518, 3.928]	101	1975-1991	Mixed	edu_basic	25_44	False	(23800.0, 43850.0]	two_parent_family	apartment middle
(12.734, 15.091]	105	0-1945	Mixed	edu_basic	25_44	False	(23800.0, 43850.0]	two_parent_family	apartment middle
(15.091, 19.044]	105	0-1945	Mixed	edu_basic	25_44	False	(23800.0, 43850.0]	two_parent_family	apartment middle
(19.044, 25.426]	106	0-1945	Mixed	edu_basic	45_64	False	(23800.0, 43850.0]	two_parent_family	terraced house
(25.426, 310.16]	106	0-1945	Mixed	edu_basic	45_64	False	(23800.0, 43850.0]	two_parent_family	terraced house
(3.928, 6.284]	109	1975-1991	Mixed	edu_basic	25_44	False	(23800.0, 43850.0]	two_parent_family	apartment middle
(6.284, 8.876]	106	0-1945	Mixed	edu_basic	25_44	False	(23800.0, 43850.0]	two_parent_family	apartment middle
(8.876, 12.734]	105	0-1945	Mixed	edu_basic	25_44	False	(23800.0, 43850.0]	two_parent_family	apartment middle

Table 8 Sensitivity analysis if of the socio-economic variables in the decision tree model

While still not ideal, we have used most representative statistical personas, i.e. a most frequent personal profile, to describe the consumption group. To visualise the commonalities we have clustered serialized attributes.

Two highest consumption groups were highly similar, the second and third group slightly less and the lowest consumption group was most dissimilar pointing to the existence of the independent third group of people.

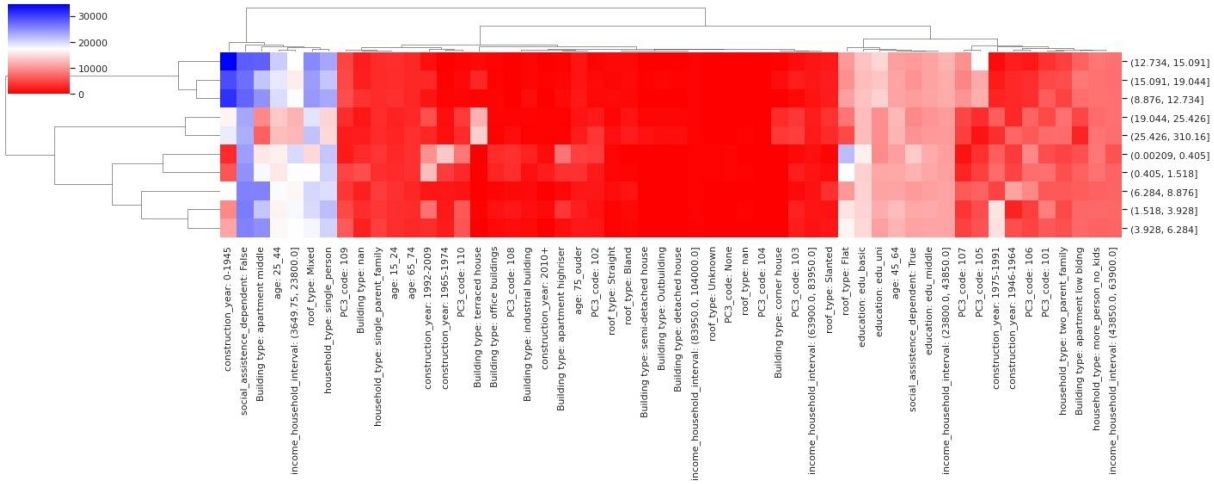


Figure 28 Clustered correlations of the most representative statistical personas in each consumption group

The ten quantiles seem to be fine grained enough to cluster based on common features. The features point to realistic **energy classes A: 0-1.5, B: 1.5-8.8, C: 8.8-19, D: 19-310m^3 p.a.**

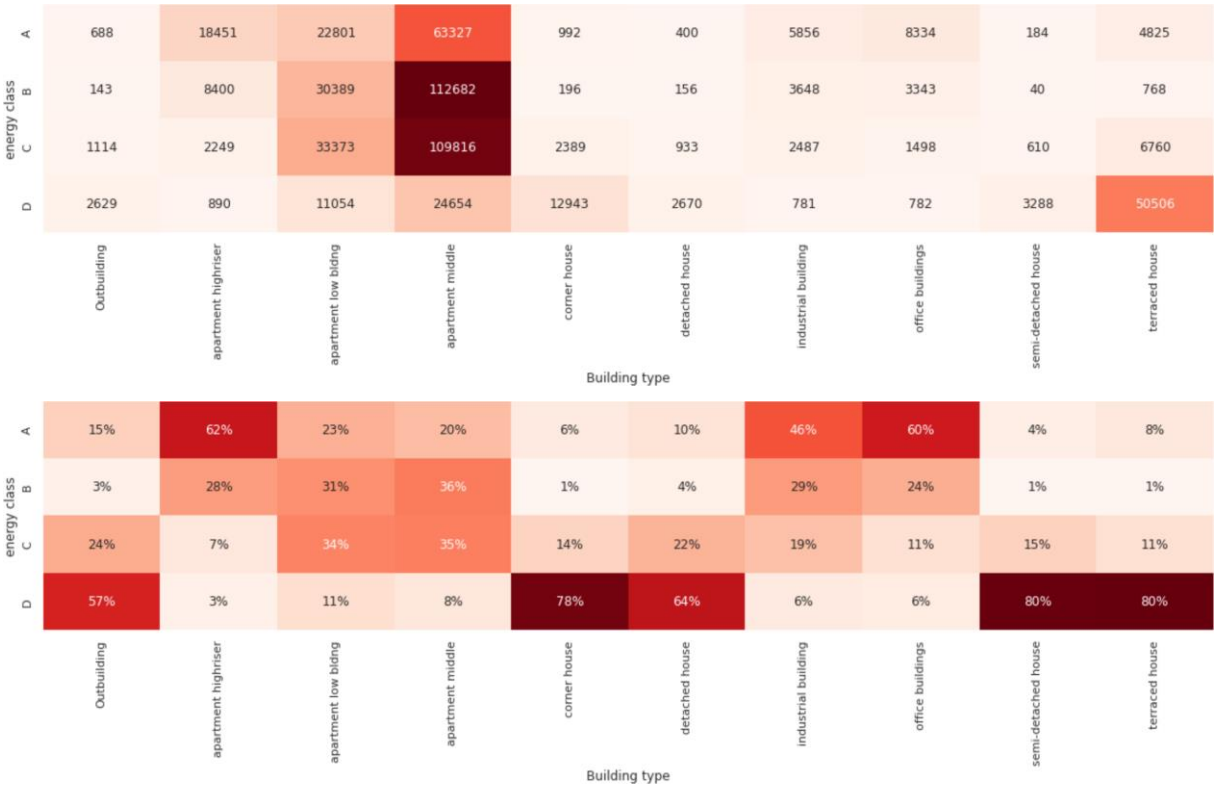


Figure 29 Artificial energy classes of Amsterdam buildings with absolute and relative counts by a building type

The most distinguishing features were dependence on social assistance, construction year before 1946, presence of the mixed roof type and a low/middle apartment building, single-person household, age 25-44 and low personal income.

We have pivoted the most important variables against house types and visualised the normalised per square meter gas consumption:

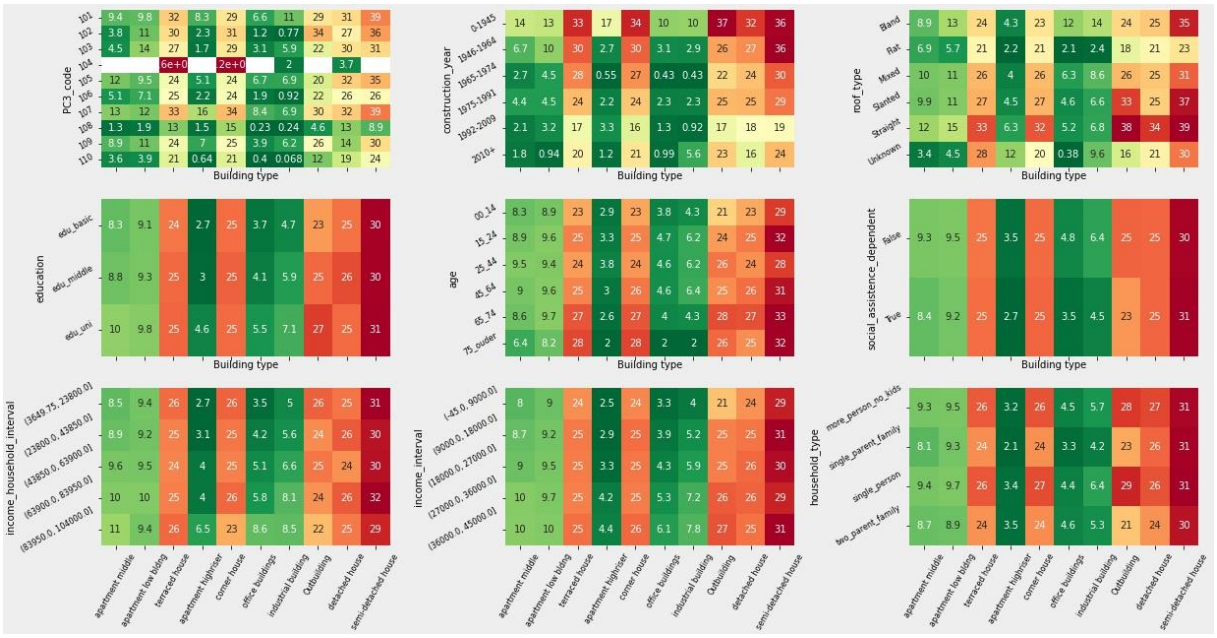


Figure 30 Pivoting individual data against house type and household consumption of gas per m^2

Synthesis of all the inputs so far lead to these key findings:

1. ***Majority of the population lives in apartments.*** *Older more consuming low apartment buildings are situated mostly in the city centre. Newer less consuming mid-level buildings usually consume on average 30% less gas per square meter. Additional almost 30% less gas consumption enjoy people in the new high-risers.*
2. *Only 3.9% of houses (and 1.4% of the population) are (semi-)independently standing, and inhabited usually by high-income households, i.e. those who can easily afford to invest in energy transition without this co-funding scheme. We expect that households with incomes lower than 20.000 EUR per annum are, on the contrary, less susceptible to make use of the loan instrument, widening further the virtual scissors in energy poverty.*
3. *The* ***"eat or heat"*** [28] ***dilemma*** *was almost palpable in the population dependent on state subsidies (elderly, unemployed, on maternity leave, …). Their higher need for heating hours due to their extended stay at home was contradicted by the average ~20% lower gas consumption.*
4. ***Single-parent families*** *and* ***75+ age group*** *consume the least energy per capita in most housing environments.*
5. *Developed energy classes point to realistic distributions of households by groups of energy consumed. Terraced, detached, semi-detached houses and corner houses are present usually in the D class, only detached houses seem to have many houses in the A and C classes too.*
6. *The ethos of educated population being more environment conscious was not confirmed,* ***higher education levels lead to higher gas consumption*** *in most house typologies. The household income and personal income correlate with higher per meter gas consumption. Higher education leads to higher wages. While this is may be the most possible explanation, it changes nothing on the fact that there is a decoupling of the education/income versus clima-related behaviour. There is obviously space for introduction of policy instruments, either voluntary or educational.*
7. *Ongoing energy transition from gas to electricity and communal heating were the most important unaccounted contributions that may affected the outcomes of this analysis. There were no direct nor proxy data available for rectification.*

5.4 Lessons learnt

This analysis has been performed for research purposes only and would require extensive crosschecks and consultation with stakeholders to be truly valid.

Our step-wise data exploration shown how much knowledge can be gained from different aggregations:

1. *The summary data gave us misleading outcomes. While this is the standard and fastest approach to policy-support analysis corroborated by the official numbers, there is a disconnection of these numbers from the local situation. Independent verification and scrutiny from public, academia or other institutions may be difficult.*
2. *The official census statistics provided distinct value in each variable without preserving any relations between variables. We still managed to extract some weak signals from these data that pointed where to dig deeper.*
3. *The population synthesis was burdened by utilisation of upstream distributions learnt from data frequencies. While administration of the city of Amsterdam has all the relevant detailed data, we could only approximate. Yet the stochastic analysis based on the synthetic population uncovered signals that seems rather strong and robust.*

We have demonstrated how difficult can be the justification of a policy instrument designed upon aggregated numbers, discovering that the loan for energy efficiency measures is not the best instrument to achieve energy transition targets. Richer population will welcome any savings and the number of houses that will be made more efficient will grow by a little, disproportional to the population in need.

[28] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1447925/

None of the tested instruments from the complete policy mix has been designed for the lower income population, which cannot afford the energy loan, and for the elderly population that has little motivation to actually invest into the efficiency measures. Many Amsterdam inhabitants live in social housing and an independent study would be needed to assess the extent of the population that has potential to actually implement energy-saving measures.

This type of demographics targeted policy instruments calls for **iterative policy co-design**. We have provided inputs to the first discussion between the politicians, the inhabitants of the houses in need of insulation, and the experts. The outcomes of this consultation would be a precious input for analysts to complement the data driven insights we illustrated in this section.

In the context of this section, we have not utilized the full scope of data at our disposal: indeed, beyond what we delved into so far, we have run corroborative studies such as a survey on mapping reaction of social housing tenants to introduction of energy saving measures in collaboration with Housing Europe and a twitter stream analysis to understand reactions and emotions of Amsterdam citizens to introduction of the energy transition policies. The results will be published in follow-up reports. To fully grasp the situation as the input to policy instrument co-design, based on all the inputs we have collected so far we would need to design and run surveys and polls to study the underlying fabric of people's well-being, etc. But by doing this, we would not only hand over a well-designed instrument, but also measured acceptance and frictions of such a disruptor on the receiving side, a real fuel for politicians.

Our goal here was to show that sociodemographic barriers to policy implementation can be identified during the ex-ante policy design phase. For a policy support analyst, this approach opens a range of opportunities where the policy would meet consenting clients instead of population in dissent that must be bent to the rule. Our analysis was more forensic than design-lead as the most important element missing was the sociologist at the start who would design the sociodemographic archetypes that the analyst may use to design a robust mix of qualitative-quantitative methods.

We may live on the verge of times for a new policy design where: an expert identifies community with all their problems, a politician who has a target to talk to and who provides feedback to a policy maker, who links issues with the human archetypes.

6 Use case 3: Generation of synthetic patient records using generative adversarial neural networks

The research community is increasingly requiring the sharing of data, in order to ensure the reproducibility of methods, analyses and results, and this includes health data (Taichman et al, 2017). The potential advantages of health data sharing entail genetic studies, cancer/chronic disease registries, substance abuse, population health management, larger-scale analytics, epidemiology/disease tracking, and even interoperability for routine patient care in the emergency department[29]. However, it is key to ensure that such data, when released, do not leak personal information.

Patient cancer data are subject to extensive privacy issues. Re-identification of a real patient can have a devastating effect both for the patient and for the trust in the relevant data-registration process.

Yet the information in the patient records have enormous value and should be made available for research and policy advice.

Classical privacy protection methods fail on large data sets[30]. Anonymization techniques (data masking, obfuscating, id removal) destroy most of the valuable information in the datasets, which significantly reduces their utility for research.

Moreover, in the era of big data, classic anonymization techniques fail to protect against de-anonymization. Researchers have demonstrated repeatedly how easy it is to re-identify data subjects in these supposedly anonymous datasets. For example, 80% of credit card owners can be re-identified by only 3 transactions[31]. 87% of all people can be re-identified, merely by their date-of-birth, their gender and their ZIP code of residence (Near & Abuah, 2021). Thus, relying on these outdated techniques puts organisations at regulatory, reputational, and financial risk.

Deep learning methods have become the state of the art in data synthesis. Advances in machine learning enable the generation of highly realistic and highly representative synthetic datasets that resemble the characteristics as well as diversity of real data sets. Synthetic data are capable of retaining up to 99% of the value and information of the original data, thereby allowing them as a proxy for research purposes.

The pitfall of classic anonymization techniques is that they mask only parts of the data while leaving everything else intact. In the era of big data, there is no non-sensitive attribute – and leaving information intact provides a target for de-anonymisation attacks.

Synthesising data, on the other hand, is a fundamentally different approach to big data anonymisation. Instead of changing an existing dataset, a deep neural network automatically learns all the structures and patterns in the actual data. Once this training is completed, the model leverages the acquired knowledge to generate new synthetic data from scratch. This artificially generated data is highly representative, yet completely anonymous. As it does not contain any one-to-one relationships to actual data subjects, the risk of re-identification is effectively eliminated.

6.1 Data provided and processing

The study worked on an anonymised data sample of 484,333 records on 456,919 patients combining data from undisclosed cancer population registries accessible by the JRC. Data were converted from wide format (patient and cancer case in the same record) to hierarchical format (table of cancer cases linked to a table of patients). The data needed extensive pre-processing as most of the fields would have been meaningless with regard to training a neural network (e.g. month). Dates were created from the year/month/day fields, incidence was translated to age as the meaning is the age of incidence. An important semantic obstacle was the simultaneous use of the ICD-9 vs ICD-10 dictionaries which required use of the ICD crosswalk[32].

[29] https://healthitsecurity.com/features/benefits-challenges-of-secure-healthcare-data-sharing

[30] http://latanyasweeney.org/work/identifiability.html , https://en.wikipedia.org/wiki/AOL_search_log_release , https://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf , https://www.nature.com/articles/s41467-019-10933-3 etc.

[31] https://epicproject.eu/fileadmin/user_upload/Mitchell_Bradly_Singapore_Mostly_AI_-_Synthetic_Data_Engine.pdf

[32] https://seer.cancer.gov/tools/conversion/2014/ICD9CM_to_ICD10_2014CF.xls

6.2 Data synthesis using the open source Synthetic Data Vault (SDV)

It is possible to generate synthetic data for different data modalities, including single table, multi-table and time-series data using mature algorithms and models (e.g. TVAE, CTGAN, Gaussian copulas, or combinations of those) together with robust benchmarking methods. It is possible to synthesise tables where the original has primary keys, fields requiring anonymisation (e.g. address), number of different data types - categorical, numerical, discrete-ordinal and date-times. Data can be generated using a number of constraints – unique combinations, value thresholds, or custom formulas such as binning, or hierarchical models. Our methods to generate time series include the **probabilistic auto regressive model (PAR)** (Percival, 1993) that allows learning multi-type, multivariate time-series data and then generates new synthetic data with the same format and properties as the learned one.

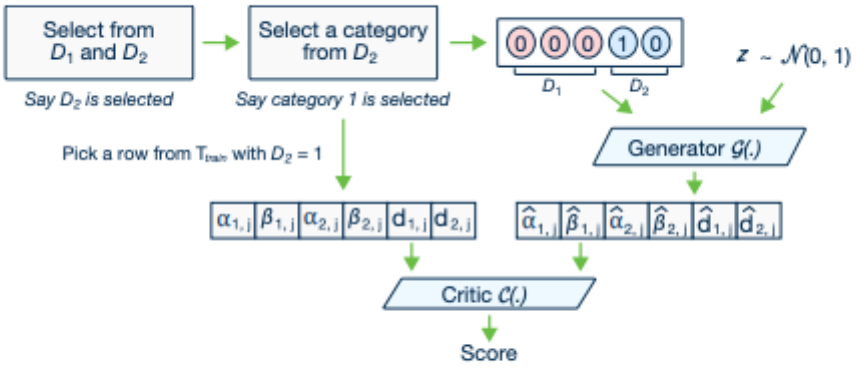


Figure 2: CTGAN model. The conditional generator can generate synthetic rows conditioned on one of the discrete columns. With training-by-sampling, the *cond* and training data are sampled according to the log-frequency of each category, thus CTGAN can evenly explore all possible discrete values.

Figure 31 Workflow of the CTGAN model. Image taken from: https://www.maskaravivek.com/post/ctgan-tabular-synthetic-data-generation/

For generation of the synthetic data from tabular data using **generative adversarial networks (GANs)** (Goodfellow, 2016), using in particular the CTGAN model (Xu, 2019), we had to use an hypertransformer[33] to convert the variables to floating point numbers ingestible by the neural network. We introduced more than 1.500 arbitrary manually designed constraints to circumvent non-zero probabilities being generated for impossible events such as prostate cancer and female gender at birth (which are mutually exclusive). The manpower invested in data conversion and cleaning exceeded two person-months.

More details on tabular GANs work can be found at the URL: https://towardsdatascience.com/how-to-generate-tabular-data-using-ctgans-9386e45836a6 and there is an excellent overview of the available methods at: https://towardsdatascience.com/review-of-gans-for-tabular-data-a30a2199342.

SDV is meant primarily for a single table synthesis. It sports the HMA1[34] hierarchical table structure synthesis method too, but only using Gaussian distributions. Since the dates in the provided data follow Gumbel distributions (age, date of birth) as shown in Figure 32, the HMA1 methods failed to train correctly.

[33] https://github.com/sdv-dev/RDT
[34] https://sdv.dev/SDV/user_guides/relational/hma1.html

	sumsquare_error	aic	bic	kl_div
gumbel_l	0.000030	1206.384136	-1.128179e+07	0.002495
logistic	0.000680	1168.011815	-9.777301e+06	0.075173
dweibull	0.000717	1171.466426	-9.751812e+06	0.078985
hypsecant	0.000739	1148.755499	-9.737515e+06	0.084872
dgamma	0.000799	1156.164247	-9.700313e+06	0.082554

	sumsquare_error	aic	bic	kl_div
gumbel_r	0.001866	1370.340823	-8.820578e+06	0.062337
logistic	0.002095	1221.965416	-8.773008e+06	0.091853
norm	0.002191	1310.277767	-8.752642e+06	0.086809
laplace	0.002567	1176.335123	-8.662701e+06	0.130863
gumbel_l	0.004105	1530.630612	-8.465734e+06	0.318424

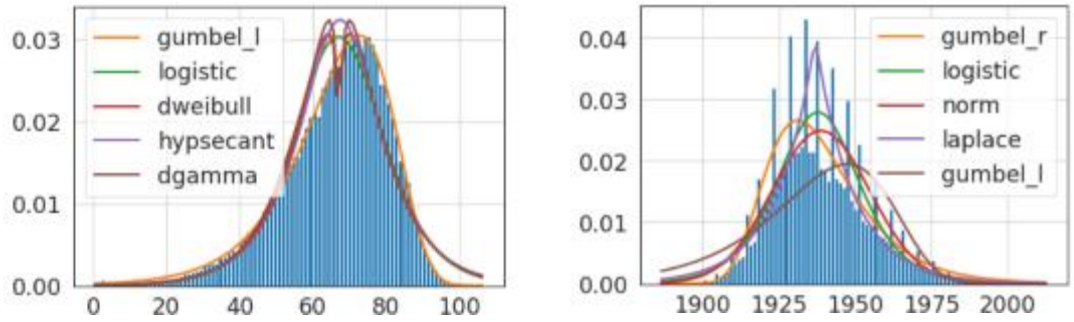


Figure 32 Distribution of age (left) and date of birth (right) in the dataset.

Exacerbated by certain problems (e.g. hard break of data submission introduced in the field last known vital status), the resulting distributions from the HMA1 synthesis are shown in Figure 33 (red – last known vital status, green – synthetic, blue – date of birth, black – synthetic date of birth):

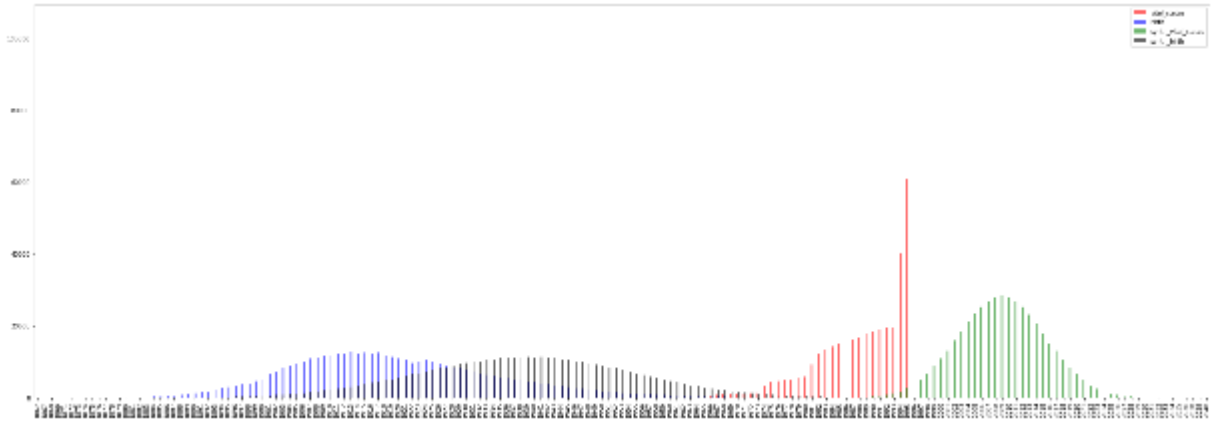


Figure 33 Distribution of HMA1 synthesis. Red: last known vital status. Green: synthetic last known vital status. Blue: date of birth. Black: synthetic date of birth.

The CTGAN and TVAE methods were successfully trained on the patient table but it was impossible to correctly train the cancer table due to non-trivial link between the patient and the cancers.

The system is available in-house and has been used on different data and can be further explored in future experiments on more consistent input data.

6.3 Data synthesis using MOSTLY.AI software

A robust commercial solution for hierarchical data synthesis was offered by the company MOSTLY.AI to test the synthesis quality. The results, available in the accompanying archive both as a PDF report and CSV files with the data correlations, are impressive. The generated report is available in Annex B.

The company website can be found at the URL: https://mostly.ai/ and the link to the published methodology is: https://www.frontiersin.org/articles/10.3389/fdata.2021.679939/full

Data were trained over 43 epochs for patients (training cycles in each of those the whole dataset has been consumed by the neural network) (epoch 38 was selected because it has the lowest validation loss), and 66 epochs for the cancer table:

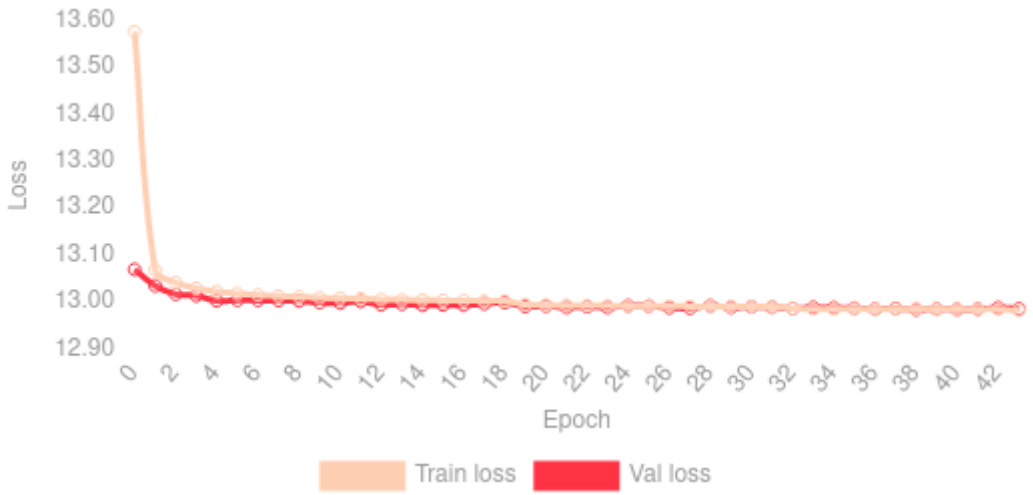


Figure 34 Loss function during the training

Data were sampled to 1,500,000 records to allow better exploration of distributions and descriptive statistics.

MOSTLY·AI

Executive Summary

Overall Accuracy: 98.8% ✓ Privacy Tests: ✓

Table	Columns	Target Data	Synthetic Data	Accuracy
patient	4 categorical 2 datetime	456,919 rows	1,500,000 rows	univariate: 99.3% bivariate: 98.9%
cancer_split_TMBG	3 numeric 5 categorical	479,883 rows	1,564,450 rows	univariate: 99.3% bivariate: 98.5% autocorrelation: 99.2%

Table 9 Executive summary table for the training with MOSTLY.AI

Key findings not described in the generated attached PDF report:

Univariate:

Values with cardinality lower than 5 were encoded as “RARE”. Topography: 106, Morphology: 633, TNM: 1009 out of 1500000 synthetized records. The more data is provided in the future, the better will the value be preserved.

Even complicated data with non-standard distributions such as the last known vital status were encoded remarkably well:

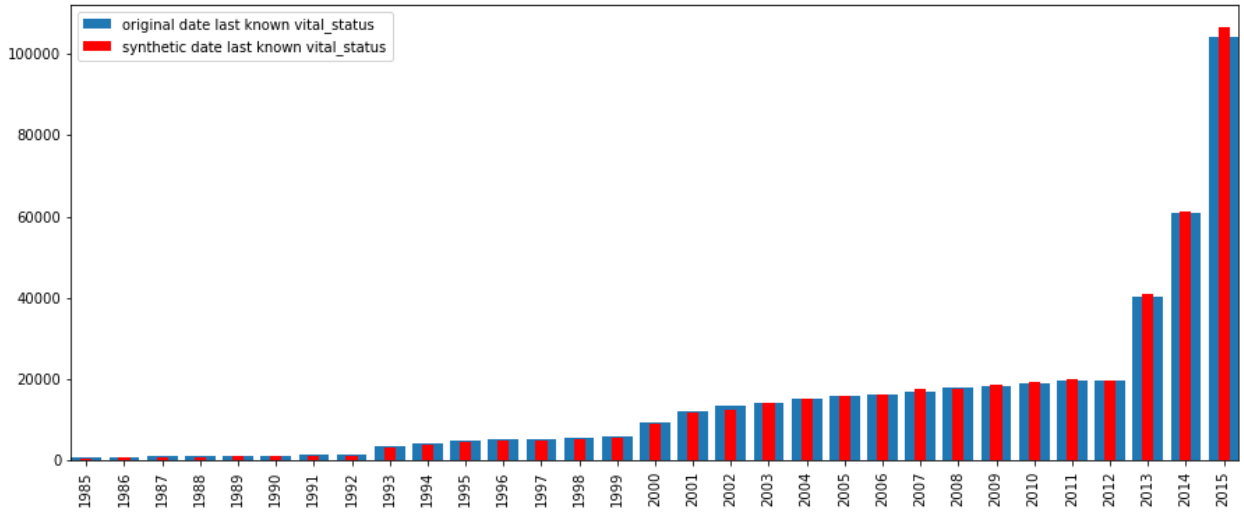


Figure 35 Last known vital status distribution in original vs. synthetic data

Bivariate distributions:

Data were checked for combinations not observed in the source data to understand how well the system has learnt the logic in data.

As we can see in the table below for prostate cancer, only 67 morphology values were not found in the original dataset:

	Toponomy	Morphology	in orig	0
10	C619	8140	True	45757
0	C619	8000	True	6967
2	C619	8010	True	868
29	C619	8550	True	242
26	C619	8481	True	89
25	C619	8480	True	59
4	C619	8020	True	22
13	C619	8201	True	21
18	C619	8246	True	20
30	C619	8551	True	19
7	C619	8041	True	18
23	C619	8350	True	14
15	C619	8211	True	13
5	C619	8021	True	13
1	C619	8001	True	10
27	C619	8490	True	9
9	C619	8070	True	7
3	C619	8013	True	4
22	C619	8310	True	2
11	C619	8141	True	1
28	C619	8500	True	1
8	C619	8050	True	1
6	C619	8033	True	1

	Toponomy	Morphology	in orig	0
21	C619	8263	False	19
14	C619	8210	False	16
19	C619	8260	False	12
20	C619	8261	False	9
17	C619	8240	False	3
33	C619	9699	False	3
12	C619	8144	False	2
24	C619	8430	False	1
16	C619	8221	False	1
32	C619	9591	False	1

Table 10 Morphology values for prostate cancer in synthetic data vs. original data

The patient records contain only a few quantitative variables – dates of birth, last known vital status, and incidence. An attempt at plotting is shown below, where negative values mean patient:

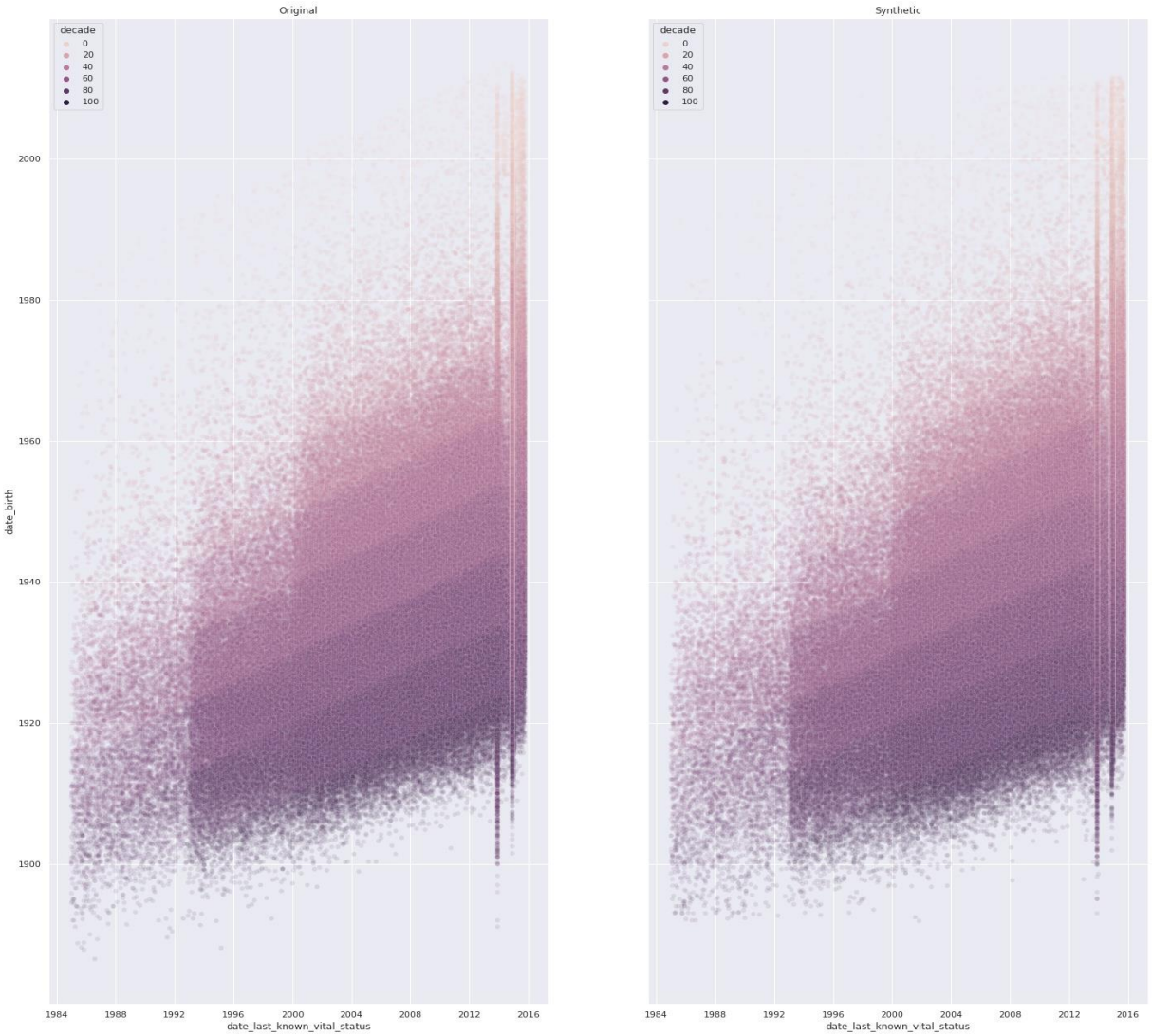


Figure 36 Quantitative variables (date of birth, last known vital status, incidence) in original dataset (left) and synthetic dataset (right).

We can check the consistence of the synthetic data by plotting the same graph but with a different colour scale for the deceased patients:

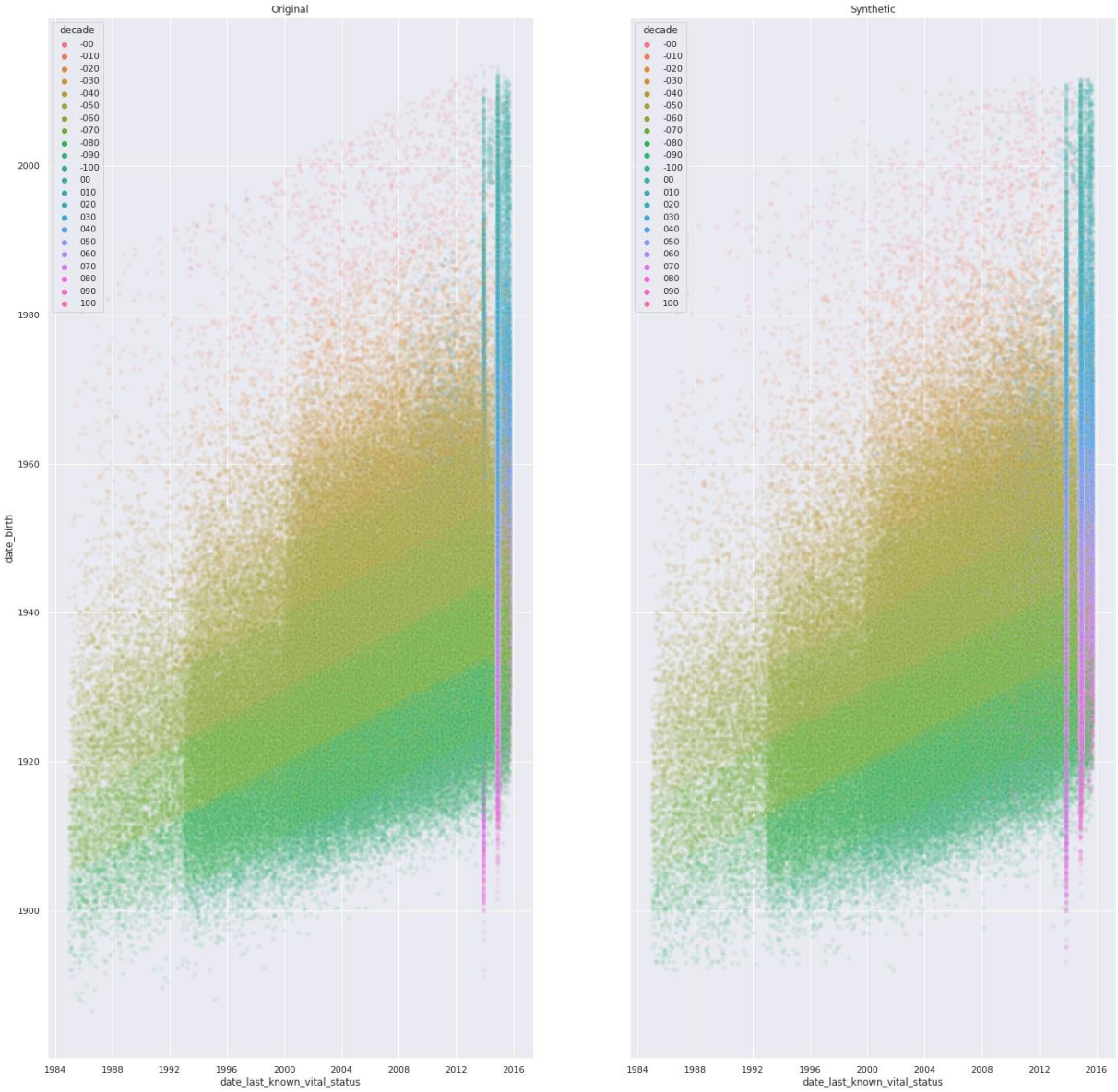


Figure 37 Date of birth, last known vital status, incidence, in original dataset (left) and synthetic dataset (right) wit deceased patients in a different colour scale.

We can see the patients who are still alive concentrated on the date line of data collection/submission. Synthetic data seem to be practically identical, perhaps with more live patients on the earlier submission date.

Another way to demonstrate consistence was elementary descriptive statistics, here for age distribution by cancer groups:

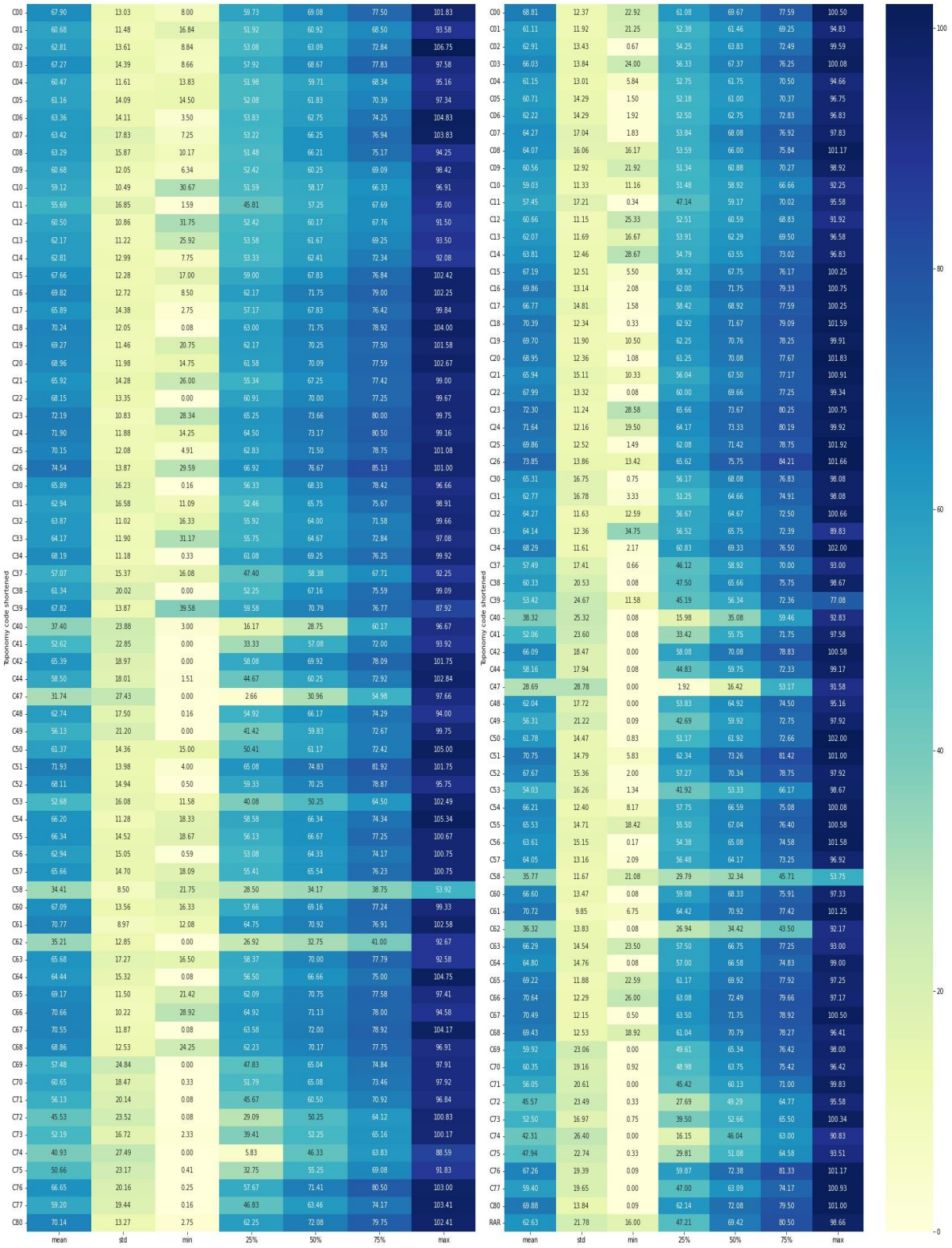


Figure 38 Age distribution by cancer groups, in original (left) and synthetic (right) dataset.

Key to the usefulness of cancer data for policymaking is the production of age-adjusted rates[35]. Comparing the original and synthetic data reveals additional information both on the original dataset and on the limitations of the synthesis process.

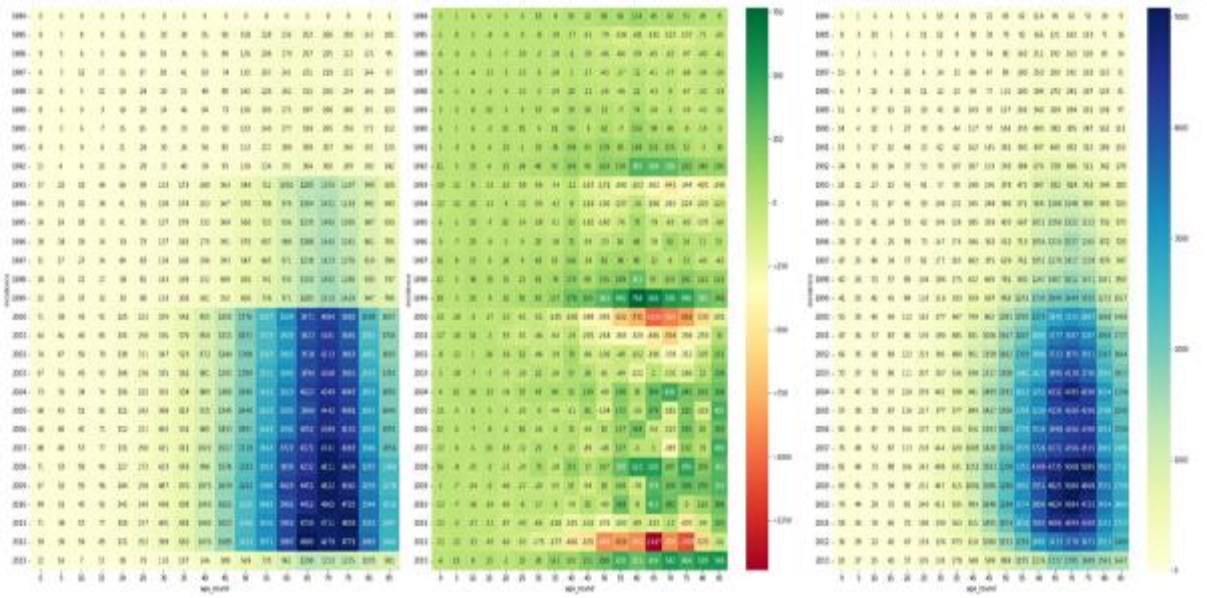


Figure 39 Incidence of cases by age class in original (left) and synthetic (right) dataset, and difference matrix (middle).

This chart shows that the original data were sampled from at least three registries, with different temporal and population coverage:

Table 11 Different temporal and population coverage of the registries

Register	Population covered	Data collection since	Latest data collection
"register 1"	Small	1984	Possibly 2013
"register 2"	Medium	1993	2013
"register 3"	Large	2000	2012

The neural network did its best to learn this discontinuity in distribution as best demonstrated in the middle image created by simply subtracting the synthetic data from the original ones. There are nicely visible hotspots on the edges where the continuous probability distribution had to fight the edge in the distribution of the original dataset.

[35] https://seer.cancer.gov/seerstat/tutorials/aarates/definition.html

All cancer sites distribution by number of cases after the ICD codes were transwalked to site names where we can see a very nice fit:

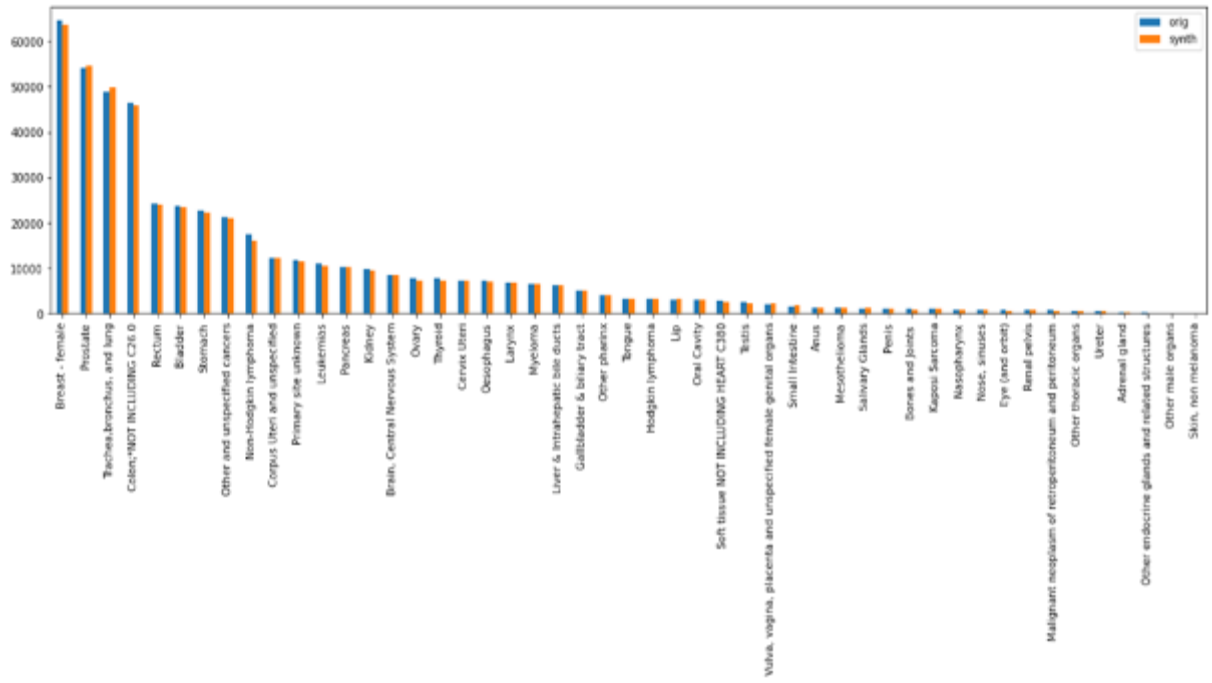


Figure 40 Number of cancer cases by site.

The capabilities of the synthesis on key cancers can be appreciated in the following chart:

Trachea, bronchus, and lung cancers, total 48761 cases in the original data, 49962 in the synthetic ones:

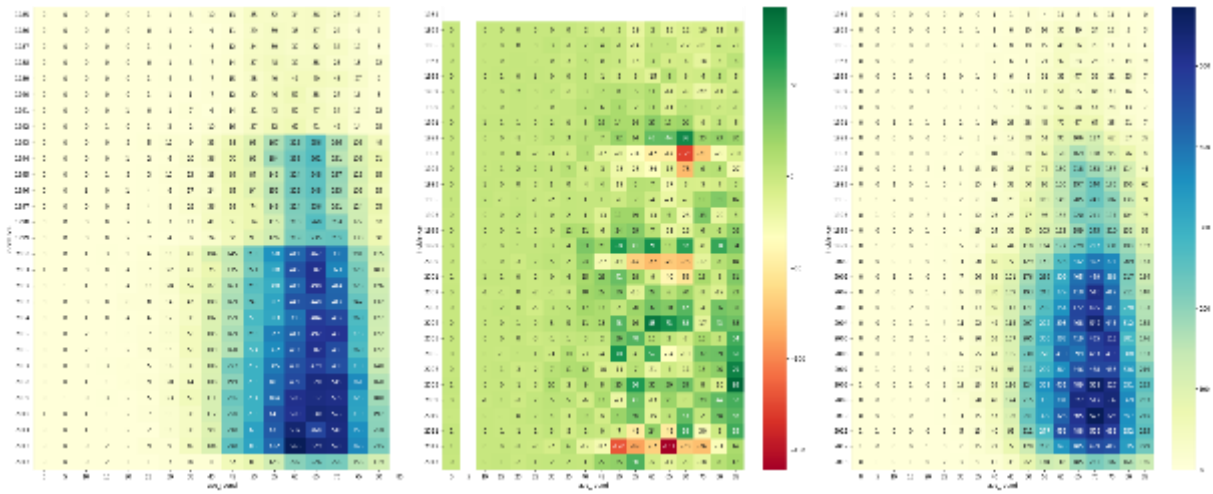


Figure 41 Incidence of cases of trachea, bronchus and lung cancers, by age class in original (left) and synthetic (right) dataset, and difference matrix (middle).

The visible shape of incidence age moving to higher age groups in “register 2” which is convincingly replicated in the synthetic data as well as the distribution shape in “register 3”.

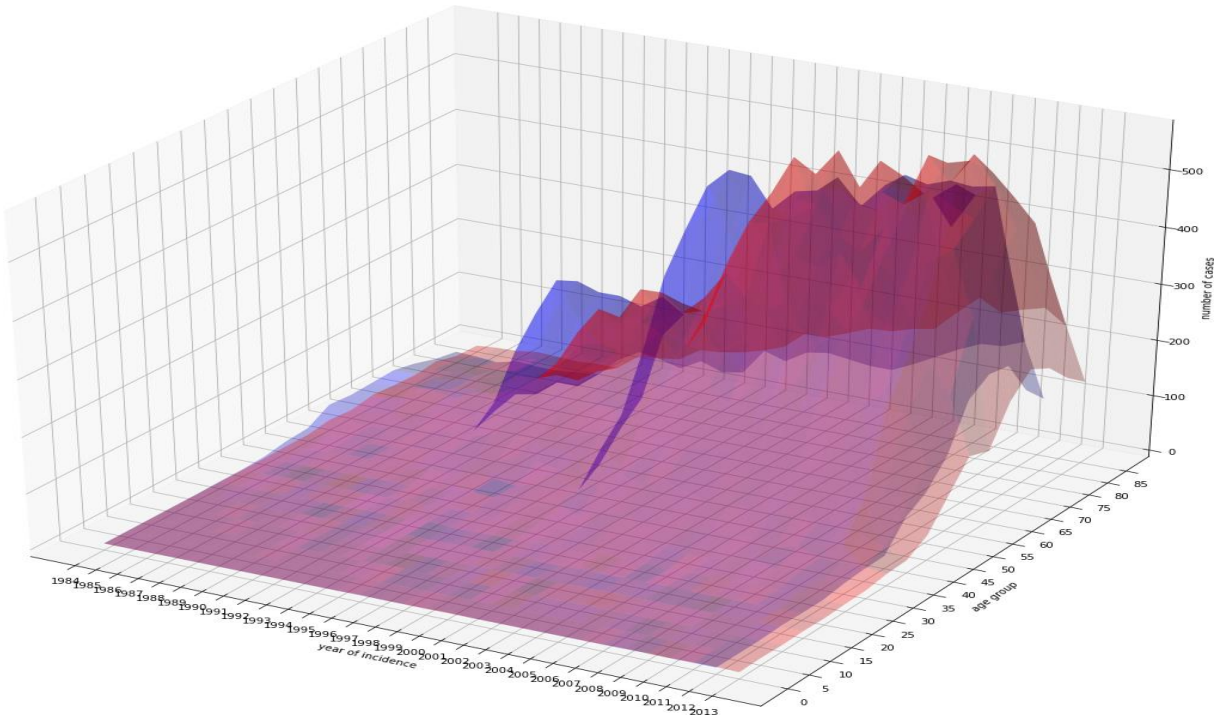


Figure 42 3D plot of the original (blue) and synthetic (red) dataset.

3D plot of the original (blue) and synthetic (red) data shows where the original dataset has more cases in higher age groups prior to 2000 while the synthetic data peak where the data from “register 3” come (cell more blue – more cases from the original data, opposite if more red). Behaviour of continuous distributional function over discontinuity of the original data play the major role here.

Stomach cancer, total 22712 cases in the original data, 22163 in the synthetic ones:

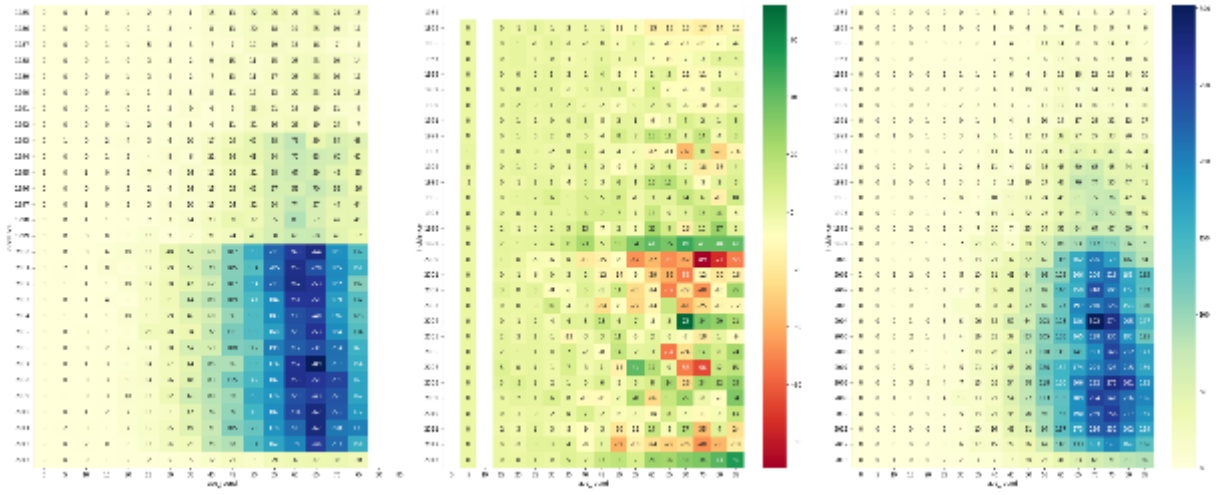


Figure 43 Incidence of cases of stomach cancer by age class in original (left) and synthetic (right) dataset, and difference matrix (middle).

Here the synthesis worked less well, yet the key information from the data is still visible – both as the shape in the pre-2000 data, and 2000-20013 diagonal development shifting again the incidence towards the higher age groups.

Tongue cancer, total 3275 cases in the original data, 3273 in the synthetic ones (0.6% of all cases):

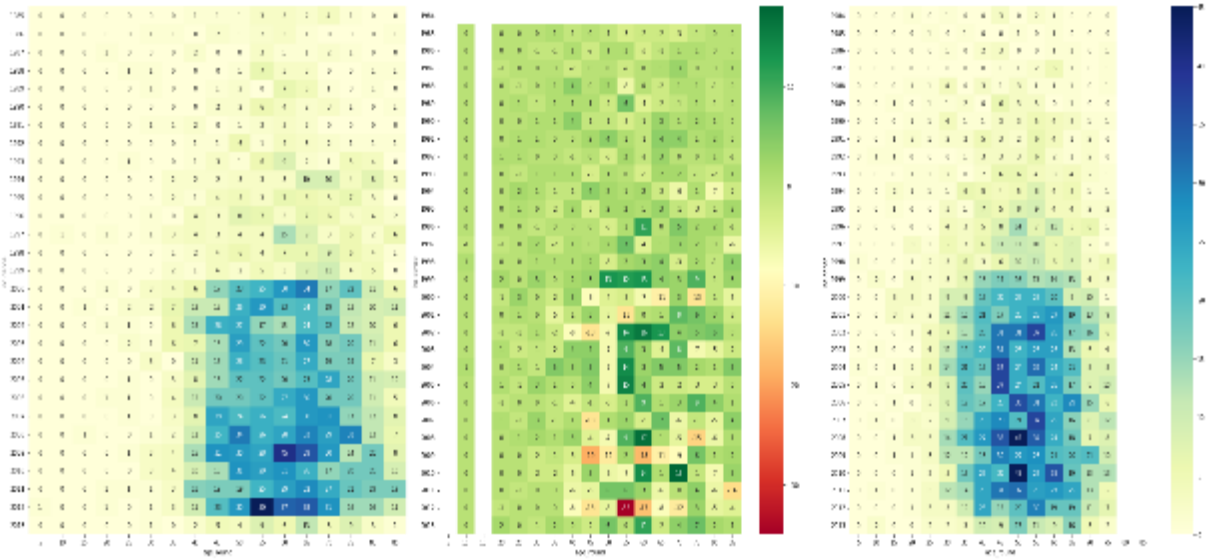


Figure 44 Incidence of cases of tongue cancer by age class in original (left) and synthetic (right) dataset, and difference matrix (middle).

Though we are already pushing the limits of what a neural network can learn, we still see a surprisingly good fit. As data are becoming sparse, the difference in specific years/age group can be higher than expected.

Renal pelvis, total 809 cases in the original data, 854 in the synthetic ones (0.16% cases):

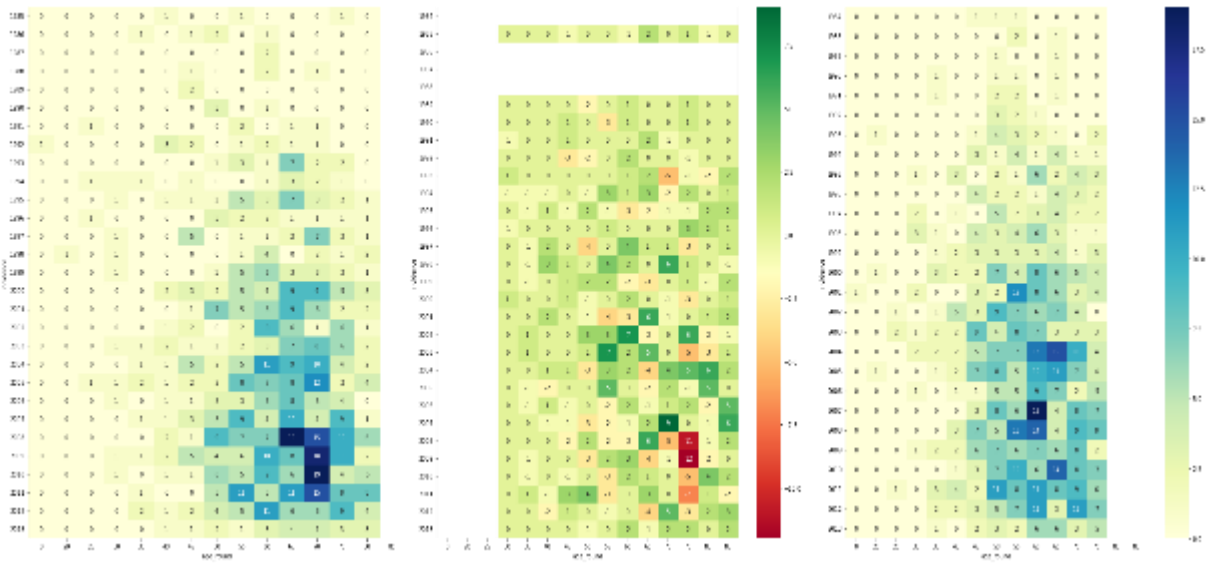


Figure 45 Incidence of cases of renal pelvis cancer by age class in original (left) and synthetic (right) dataset, and difference matrix (middle).

Unexpectedly, the overall data still show rather convincing distributions.

Adrenal gland, total 381 cases in the original data, 331 in the synthetic ones (0.06% cases):

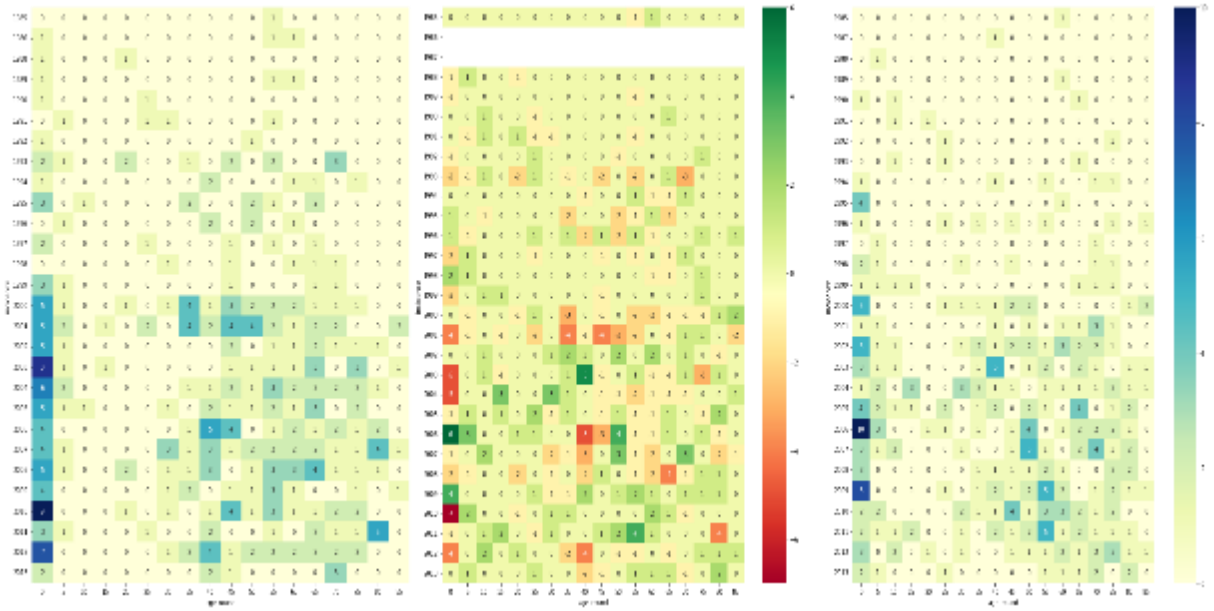


Figure 46 Incidence of cases of adrenal gland cancer by age class in original (left) and synthetic (right) dataset, and difference matrix (middle).

Other male organs, total 126 cases in the original data, 113 in the synthetic ones (0.02% cases):

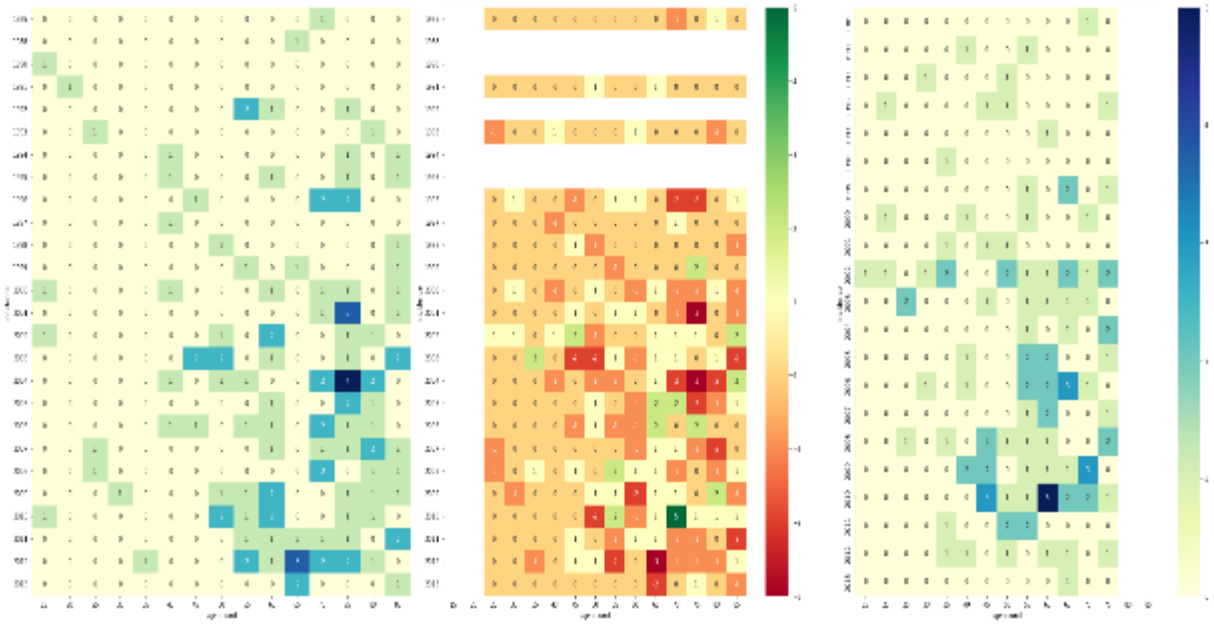


Figure 47 Incidence of cases cancer affecting other male organs, by age class in original (left) and synthetic (right) dataset, and difference matrix (middle).

Even such a weak signal was translated relatively correctly into the synthesis process.

6.4 Lessons learnt

Deep learning methods of data synthesis have specific advantages compared to statistical methods especially in terms of scalability. The level of privacy preservation can be fully controlled, checked and verified. The veracity of the data produced is thus very high. Nevertheless, the synthesis process is dependent on well-processed data with ideally continuous distributions and introduction of additional constrains is necessary.

Data for ingestion into deep learning systems need to be heavily pre-processed to avoid confusing the neural network. Controlling cardinality, meaningful dates or semantic consistence are the key prerequisite to quality outcomes. Yet the sheer quantity of data is the primary and defining quality of the input. In other words, clean and abundant data are the most important factor in training representative model.

Current methods of data synthesis using open source tools are relatively powerful but only for flat tables, with limited number of constraints, low cardinality categorical variables and continuous, without hard breaks. Hierarchical data still need industry-strength commercial solutions where elaborate heuristics and robust filters accompany the vanilla GANs/VAEs. The field is evolving very fast and we may expect competitive open source solutions in the near future.

Synthetic data have proven great potential and are the go-to methods ready to be deployed in real-life scenarios. Policy applications can now be researched and developed with little risk involved.

7 Conclusions

The future challenges such a climate change require better understanding of in-situ dynamics including behavioral patterns while protecting the people's privacy. This antimony can be overcome by introduction of synthetic data as described in this report.

The synthetic (population) data are a part of a huge global shift from privacy- and bias burdened information in order to feed data-hungry AI research. Leading strategy consultancy Gartner assesses that "60% of the data used for the development of AI and analytics projects will be synthetically generated"[36].

Every knowledge representation has both advantages and disadvantages, data included. There are many available processes for data synthesis that lead to highly different outcomes, from high precision but highly limited scalability in statistical learning to high consistence high scalability but black box deep learning models. There will be many new techniques researched and developed in the future as the needs are obvious and well-funded start ups in this field pop up every month. Therefore the conclusions do not focus on a specific technique but general rules of applicability of the synthetic generated data.

More important than focusing on how to synthesize data is what can we achieve with the new data available at scale, how to convince data owners to unleash their coveted data to the broadest audience, and how to accommodate this massive new ability into the policy formulation and assessment.

The most important **features and opportunities** of the synthetic data for policy-making and AI to make it implementable in policy cycle:

1. Privacy unburdened data
    (a) **Synthetic data are free of privacy issues** when well designed and quality checked. Primary (e.g. population) microdata cannot be shared. Anonymized data and data aggregates lose too much information. Pseudonymized data are prone to de-anonymization attacks.
    (b) Synthetic data **change everything from privacy to governance** and need a serious research invested in order to understand, pilot and implement them into a normal policy cycle.
    (c) Among the privacy-preservation technique studies analysed (differential privacy, data perturbation, homomorphic encryption, secure private computing infrastructure), data synthesis gave the **best price (effort)/cost ratio**.
    (d) A **complete data synthesis life cycle management** is fundamental for **trust in the generated data**. Data preparation, feature engineering, model hyper-parameters and tuning, training outcomes, feedback collection are just a few points on the checklist of a machine learning pipe line.

2. Breakthrough policy formulation and evaluation
    (a) As physical media from vinyl to streaming platforms define how artists express themselves, policy makers will have opportunity to create **radically new policy instruments** based on more detailed less biased population data.
    (b) Robust multipurpose synthetic population has a capability to take over the role of statistical aggregates in many domains of policy research, anticipation, formulation, implementation and evaluation.
    (c) Policy support can **embrace diversity** by stopping averaging out, and thus marginalisation of under-represented minorities by shifting away from aggregate-level statistics; capture the full diversity within the population, and start coping with its complexity, rather than continue ignoring it.

[36] https://blogs.gartner.com/andrew_white/2021/07/24/by-2024-60-of-the-data-used-for-the-development-of-ai-and-analytics-projects-will-be-synthetically-generated/

(d) Synthetic data in **policy making process** will enable involvement of much broader communities and stakeholders as there no longer will be the privacy and other concerns related to data sharing and reuse leading to viable path to a standard co-creation process.

(e) Same goes for **algorithmic governance**. Not only can synthetic data be used to model and predict the behavior and biases of the model, it can also be made available to scientist and communities for in-depth scrutiny.

(f) Synthetic population and population as a graph provide breakthrough capabilities for ex-ante evaluation of policies and policy instruments. It enables modelling of complex systems including **sociodemographic barriers** to policy implementation, creation of **targeted policies** and evaluation of impact of policies at the level of communities and individuals.

3. Consistent policy support

(a) Synthetic population can **advance the research in majority of models depending on population data**, and ensure comparability of the models. Computational requirements unavailable a decade ago is no longer an obstacle in utilization of data at the level of hundreds of millions of individuals. Even better, current dependence on microdata samples that are statistically representative but differ among studies, a situation that crushes comparability, can be abolished by using unified synthetic dataset e.g. for households.

(b) Large internet platforms are revolving around **behavioural data** and it raises the bar for the new policy requirements. On top of that, synthetic population profiling enables safe sandboxing environment to study future policy impacts.

(c) Conversion of static tables to probabilistic synthetic data facilitates **probabilistic enrichment** of dataset with data with known uncertainties such as surveys or behavioural profiles.

(d) Synthetic individual-level data can be augmented using **soft data** (opinions, sentiments) and/or big data to had better understand rifts in the society by **mapping communities** through their **sociodemographic profiles, needs and need satisfaction**.

(e) **Survey and policy experiment design** often struggles with understanding of structure of the specific surveyed part of society where

4. Towards practical AI in Europe

(a) The AI-centric EU policy strategy depends on making available extreme amounts of data and knowledge to AI research and implementation, which is unobtainable in the current settings. **Synthetic data can enable AI strategy**. Existing Data Spaces are still prone to privacy attacks, suffer from limited data circulation, access restrictions. Synthetic data have proven capabilities to establish this “outer” data space.

(b) Synthetic data have potential to enable **creation of realistic digital twins** including controlled assembly from data spaces and analytical data fusion to extract actionable insights.

(c) By **controlling bias** in data enables utilization of robust synthetic data for machine learning models. Robust data are the key to credibility of AI-based policies. On the other hand, synthetic data facilitates **AI and model explainability** of already trained models. Throwing huge amounts of real-like data on the model develops patterns among which the model decides.

5. Raise the data to the cloud:
    (a) Synthetic data **eliminates the bureaucratic burden** associated with gaining access to sensitive data and allows moving complex previously sensitive datasets to common cloud services and data spaces.
    (b) **Data democratization** – a new type of Open Data would emerge from the shift towards the synthetic data; empowerment of people with knowledge which is at the local level, i.e. relevant to them. New community-borne models will be trained on the new data available in similar way the natural language processing free open and shared models rocketed the text analytics by decades in just a few years.
6. **Data literacy** is today as what was computer literacy in 1990s. Synthetic data have potential to accelerate their learning path towards data-driven decision & policy making by making available data closer to the people's perspective.

Disadvantages of synthetic data:

1. Original data and synthetic data differ. The level of difference can be assessed, from univariate distributions to overall consistence yet the user should be fully aware of the limitations. Complete metadata on consistence checks is a must. Ideally there is a verification process in place which would allow running the same analysis on the original data after the algorithm/model has been developed on the synthetic data.
2. The process of data creation requires a rather high level of expertise with both domain knowledge, statistics and artificial intelligence, which usually requires a team of experts for reasonable synthetic product delivery. Validation and documentation are the key.
3. Synthesis of hierarchical or very complex data can be highly challenging. High cardinality values, too many variables, cut distributions need to be either sanitized or prevented. Hierarchical dependencies should be clean and minimal to avoid confusing the neural network. Commercial solutions still beat the available research and open source solutions by a huge margin at the time of writing.
4. Data synthesis becomes very difficult or downright impossible on small datasets. 10,000 records can be understood as a bare minimum for a small number of variables. Datasets containing tens of variables would need to start at hundreds of thousands or millions of records.
5. Quality of synthetic data is highly dependent on the quality and robustness of the model that created it. There is ongoing research on attacks on model training such as adverbial perturbations but this is common with AI and machine learning model training.
6. Synthetic data can be used for creation of data aggregates as they are only probabilistically correct at the individual record level. This feature protects privacy but limits uses of the synthetic data for specific analytical purposes.
7. It is impossible to join synthetic data with other dataset at the record level. It can be done first, and run the data synthesis on the joint datasets. Also using aggregated data means that the model would learn upstream aggregation model instead of the knowledge in data.
8. If synthetic data isn't nearly identical to a real-world data set, it can compromise the quality of decision-making that is being done based on the data especially if taken at face value. User discretion should be advised by the data provider.
9. New methods are being researched and developed continuously. The concept of synthetic data is to change perception of how the privacy-burdened data can be effectively utilized but it is not the only way, new techniques will be coming. Education and getting ready for the major shift in data and model capabilities is the key benefit from embracing the synthetic data.

Annex I. The complete French population generation method

A.1 Introduction

Synthetic populations, in the form of disaggregated individual data, represent the main input entities to several multi-agent models and micro models. Such models are used in a plethora of diverse contexts in policy design and evaluation, from economical simulations (labour market policies, tax benefit, poverty-reduction policies, multi-country microsimulations, etc.) to agricultural and environmental policies, education, health, demographic and social well-being, urban modelling, just to mention a few.

In policy decision making, the levels of complexity that come into play entail: the heterogeneity of the population and its often under represented different subgroups; the behavioural response of individuals; policy complexity and impacts on different subgroups; further complexity is added by considering the time and space dimensions. It becomes apparent that, to disentangle such a level of complexity, recurring to a model is necessary (Aaberge et al. 2014).

Micro models based on synthetic population can simulate the effects of proposed policy implementation on subgroups, as well as estimate program costs and caseload (Citro & Hanushek, 1991).

The spatial dimension is important to consider whenever the model is spatially targeted. Datasets with a spatial component are usually available at census area (or coarser level of detail). The input data to the models may be rich in contextual data related to persons or households and lacking on the spatial information, or the sample size may be too limited to be representative at a fine spatial scale, or vice-versa, data at fine-grain spatial resolution may have gaps in contextual data (Aaberge et al. 2014). Generating a synthetic population filling data gaps is challenging.

Synthetic populations are a powerful tool because they can be extremely informative without breaking the privacy of citizens, but still reflecting the complexity of the structure of the population, and the characteristics of the individuals that influence their behavioural response. A synthetic population is designed to reflect the heterogeneity of the real population, including minorities and under-represented individuals that would not be characterized considering just the general statistics of the population.

Our study aims at reconstructing a synthetic population that would serve as a baseline to be used as an input to different kinds of models. This baseline should be flexible enough to be successively enriched and updated whenever more data becomes available. The population has to carry all possible information until a model is chosen and the relevant features can be selected accordingly. The baseline is characterised by all features present in the source.

The choice of reconstructing the French population is not casual. Other Authors (Antoni et al. 2017; Delhoum et al. 2020) as well have successfully adopted French case studies, because the French INSEE published the complete data model of the census data. This latter allows for a complete reconstruction of synthetic individuals with their attributes by means of merging the different table and finally un-weighting the individuals / households, as explained hereinafter.

In the proposed approach, people are modelled at the individual level and households have been reconstructed for all the individuals, so that no choice has been performed between one representation and the other. This is possible because we do not start from a micro sample. Another trade-off always present in literature is the selection of a subset of attributes characterising the records. The selection is made before the actual population reconstruction. This is because the optimization techniques require choosing which parameters we require more fitting to the real population statistics, and the optimization necessarily is performed at the expense of the other parameters, for which we accept a certain relaxation.

In our work, we preserved all the features available for each record. The selection of the features is operated only whenever the population is used as an input in a model, in a particular use case. From the complete synthetic population is also possible to extract aggregated statistics for various combinations of attributes. We structured the population as a georeferenced database that can be served to the user in the form required by the user's model, as a database view. It is possible to create customized queries that regroup the records of the database according to the characteristics of interest that can be related to socio-demographic, education, health, work conditions, etc.

In many studies available in literature, artificial populations have been reconstructed for a particular area and for a particular case study (Farooq et al. 2013; Ye et al. 2009; Antoni et al. 2017; Lenormand & Deffuant, 2012; Gargiulo et al. 2010; Delhoum et al. 2020; Thiriot & Sevenet, 2020; Namazi-Rad et al. 2014). In our work, we reconstructed the population of the entire France at the individual level, and at

the same time we reconstructed the households, without discarding any characteristics given in the initial data. We merged data from different sources, from census to OpenStreetMap, eventually placing the households into houses, individuating working places and assigning them to each working individual, individuating commuting routes and assigning to each individual the most probable Points Of Interests (POIs), such as visited commercial activities, leisure and sport facilities, nearest health services, educational institutions, etc.

A.2 Data sets description and preparation

The main data sets used in this study are provided by the French National Institute of Statistics and Economic Studies INSEE (Institut National de la Statistique et des Études Économiques)[37], and refer to census data collected in 2016[38].

INSEE collects, produces, analyses and provides information about the French economy and society. The anonymized detailed files provided by INSEE aim to provide informed professional users (public bodies, local authorities, large companies, consultancy and consultancy firms, researchers, etc.) the opportunity for personalized use of population census data. These files make it possible to carry out exploratory analyses of data, to model behaviours, or simply to tabulate on a particular subpopulation defined according to certain criteria: belonging to a geographical area and / or statistical unit presenting certain characteristics.

A.2.1 Definitions

The official documentation of INSEE statistics[39] presents definitions useful to understand our work:

"[…] **Associated or delegated municipalities:** In application of the law n ° 71-588 of July 16, 1971 on the regrouping of municipalities and of the law n ° 2010-1563 of December 16, 2010 on the reform of local authorities, a certain number of municipalities resulting from mergers have one or more "associated or delegated municipalities". The population of a municipality fraction is the municipal population calculated for this municipality fraction.

Cantonal fractions: A certain number of communes, in general the most populated, are divided into cantonal fractions. The population of a municipality fraction is the municipal population calculated for this municipality fraction.

Number of municipalities: When, in a department, the territory of a commune is distributed among several cantons, it counts as one unit in the number of communes of each of these cantons, but only counts as one unit in the number of communes of the district and department. This explains why the number of municipalities in a district (or department) is not always the total of the number of municipalities in the cantons that make it up.

Territorial limits: [...]

INSEE assigns each region, department, arrondissement, canton and municipality, a code on 2, 3, 1, 2, 3 positions respectively. A district, a canton or a municipality is perfectly identified by the concatenation of the code of the department in which it is located and its own code. A cantonal fraction is identified by the code of the canton to which it belongs and the code of the municipality.

The reminder of these different codes in all the tables makes it possible to know the cantonal and municipal composition of the arrondissements as well as the municipal composition of the cantons."

"In order to prepare for the dissemination of the 1999 population census, INSEE developed a system for dividing the country into units of equal size, known as IRIS2000. In French, IRIS is an acronym of 'aggregated units for statistical information', and the 2000 refers to the target size of 2000 residents per basic unit.

[37] INSEE. National Institute of Statistics and Economic Studies. Available online: https://insee.fr/
[38] https://www.insee.fr/fr/information/4172214
[39] https://www.insee.fr/fr/statistiques/4265429?sommaire=4265511#documentation

Since that time, IRIS (the term, which has replaced IRIS2000) has represented the fundamental unit for dissemination of infra-municipal data. These units must respect geographic and demographic criteria and have borders, which are clearly identifiable and stable in the long term.

Towns with more than 10,000 inhabitants, and a large proportion of towns with between 5,000 and 10,000 inhabitants, are divided into several IRIS units. This separation represents a division of the territory. France is composed of around 16,100 IRIS, of which 650 are in the overseas departments.

By extension, in order to cover the whole of the country, all towns not divided into IRIS units constitute IRIS units in themselves.

There are three types of IRIS unit:

The residential IRIS: population generally falls between 1,800 and 5,000. The unit is homogeneous in terms of living environment and the boundaries of the unit are based on the major dividing lines provided by the urban fabric (main roads, railways, bodies of water etc.)

The business IRIS: containing more than 1,000 employees, with at least twice as many employees as other residents.

The miscellaneous IRIS units: specific large zones, which are sparsely inhabited and have a large surface area (leisure parks, ports, forests etc.).

As of January 1st 2008, 92% of IRIS units were residential, with 5% business. Since their creation, the demographic characteristics of certain IRIS units may have evolved, although their classification will not have been updated.

In 2008, a very partial reworking of the division system was undertaken to take into account major developments in the road network or demographics. This reworking was limited to around 100 IRIS units, in order to preserve continuity in the data publication series.

Division of a territory into IRIS units may be affected by modifications in the geography of the municipalities (merging of towns or villages, founding or repopulation of municipalities, land exchanges). So it is useful to specify the year of reference, for example by noting either IRIS-geography 1999 or IRIS-geography 2008."[40]

From Wikipedia[41]:

"Each IRIS constitute a "micro-neighbourhood", made up of a set of contiguous and homogeneous blocks, bringing together 2,000 inhabitants or more. Each IRIS constitutes a basic municipal sector, a homogeneous geographic and demographic "micro-district", clearly and durably defined. This "elementary link" is used to collect statistical and demographic data. These data are then analyzed and the results published by INSEE. France has 50,800 IRIS, including 700 in the overseas departments, broken down as follows:

16,100 IRIS, including 650 in the overseas departments, resulting from the division of municipalities with more than 10,000 inhabitants, and most municipalities with 5,000 to 10,000 inhabitants,

34,800 IRIS, made up of non-divided municipalities (*communes non-découpées)*"[34].

A.2.2 The model

The data model is represented by a set of tables linked to each other by means of keys. The data are aggregated by a weight (IPONDL for LOGEMT, IPONDI for the other files).

[40] https://www.insee.fr/en/metadonnees/definition/c1523
[41] https://fr.wikipedia.org/wiki/%C3%8Elots_regroup%C3%A9s_pour_l%27information_statistique

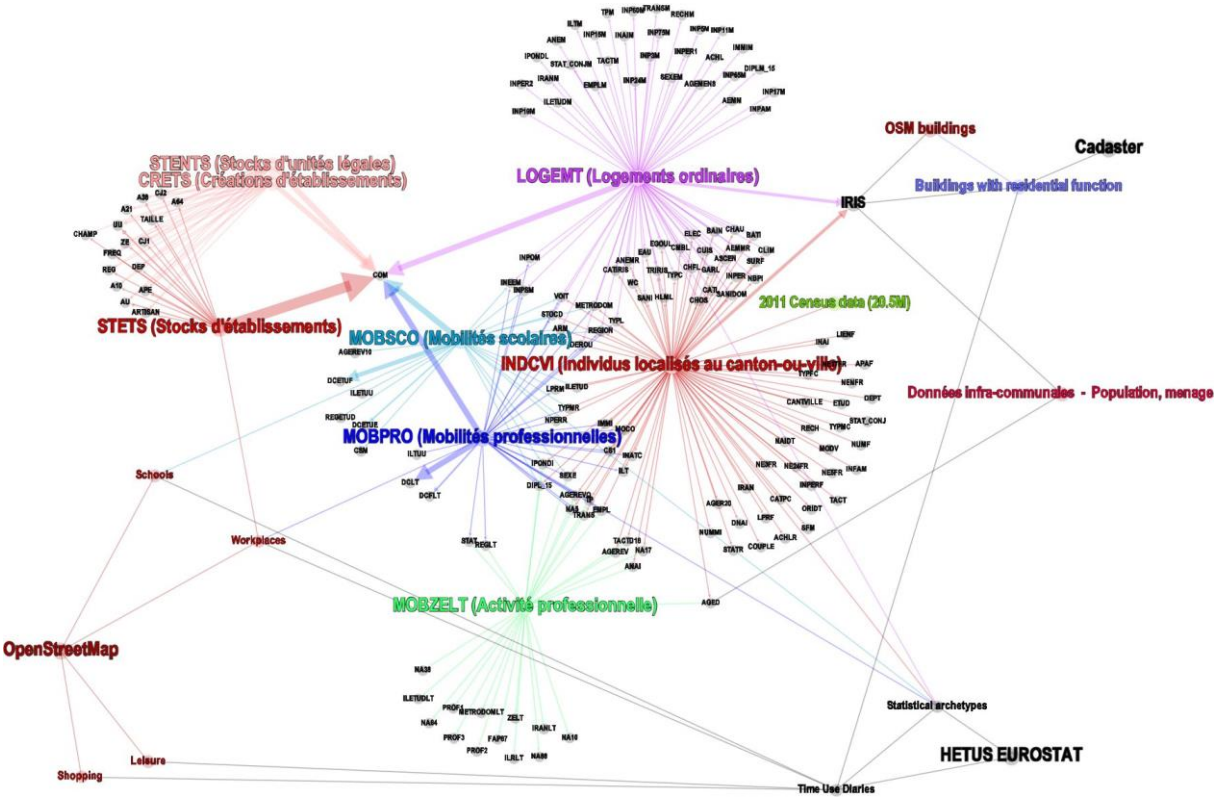


Figure 48 Representation of the data model from the INSEE.

The tables used in this study are the following:

- ***INDCVI (Individus localisés au canton-ou-ville)***

Table containing the characteristics of the individuals, such as age, sex, level of education, household composition, etc. Each record contains attributes of an individual aggregated by means of a weight (IPONDI) that gives a measure of the frequency with which every record (profile of the individual) is found in the population in a certain area. Records also contain the 5-digit code of the census area (IRIS code). Small census areas, where the privacy is at stake, are grouped into larger areas and the attributes are given at a coarser resolution.

- ***LOGEMT (Logement)***

Every record in the table corresponds to an ordinary dwelling described according to its location, its characteristics (category, type of construction, comfort, surface area, number of rooms, etc.), and the socio-demographic characteristics of the household residing there. Household information is provided only when the accommodation is occupied as the main residence. Information are aggregated by the weight.

- ***MOBPRO (Mobilités professionnelles)***

Table containing the information about professional mobility. Each record in the file corresponds to an individual described according to the characteristics of his trips to go to work (home-work trips), his main socio-demographic characteristics, as well as those of the household to which he belongs.

- ***MOBSCO (Mobilités scolaires)***

Table containing the information about the mobility related to education. Each record in the file corresponds to an individual described according to the characteristics of his trips to go to attend an education institute (home-study trips), his main socio-demographic characteristics, as well as those of the household to which he belongs.

- ***MOBZELT (Fichier Activité professionnelle des individus (localisation à la zone d'emploi du lieu de travail))***

Each record in the table corresponds to an individual located at the workplace described according to the characteristics of their trips to work (home-work trips) as well as their main socio-demographic characteristics. All active individuals with a job, aged 15 or over, registered and working in France are taken into account.

- ***BPE (Base Permanente des équipements)***

This dataset contains the localization of a wide range of facilities and services, including: public services (post office, police station, bank, court, masons, painters, electricians, plumbers, veterinaries, hair-stylists, real estates, etc.); commerce facilities (hypermarkets, supermarkets, large surface bricolage centres, grocery stores, bakeries, non-food shops such as book stores, furniture, etc.); all types and degrees of educational institutions; health facilities and services; social services; transportation infrastructures; sport, leisure and cultural facilities; touristic facilities.

- ***CONTOURS IRIS***

This dataset is provided by IGN (Institut National de l'Information Géographique et Forestière)[42]. It is a digitized background of Iris regions defined by INSEE for census.

- ***HOUSING and WORKPLACE***

The key value of the cadastral data is that they very precisely record geometric and ownership properties of buildings and plots to preserve the constitutional rights of citizens. The main issue with cadastral data for analytical purposes is the lack of interest of the cadastral agencies to maintain any other attributes.

The housing dataset was extracted from the French BD-TOPO data, maintained by the national mapping agency IGN, namely BDTOPO_3-0_TOUSTHEMES_SHP_LAMB93 files for all mainland departments and island Corse. The bâtiment (buildings) data contained in most regions information on the main and secondary function (USAGE1, USAGE2). We have performed a vast number of cross checks against other sources ranging from web maps to Google StreetView to local housing descriptions in order to understand under which conditions the residential function of the building actually corresponds to reality. In about 10% of the departments, the residential function information was not even included in the data. The data are beset by a huge number of uncleaned attributes, ranging from area of the building to attributes.

[42] https://www.data.gouv.fr/fr/datasets/contours-iris/

For instance, this is the overall statistics of all independent buildings in France with surface area smaller than 10m^2:

Château	Indifférencié	8
	Résidentiel	18
Indifférenciée	Annexe	1162
	Commercial et services	1475
	Indifférencié	22789
	Industriel	10
	Religieux	4
	Résidentiel	174005
	Sportif	3
Industriel, agricole ou commercial	Agricole	37
	Annexe	14
	Commercial et services	43
	Indifférencié	297
	Industriel	22
	Résidentiel	270
Moulin à vent	Résidentiel	1
Silo	Indifférencié	1
Tour, donjon	Indifférencié	11
	Résidentiel	8

Only 43% of all residential buildings have information about the number of flats. Many buildings have inconsistent attributes even if these are two buildings in the same town built in the same place and year, of the same size, and Google StreetView shows the full use of both buildings.

As a final step, to verify the residential function, we have performed a spatial join with OpenStreetMap data and using cKDTree from python scipy.spatial library we have assigned the IRIS and Arrondisement values to the buildings. Resulting data contain information on number and size of apartments in the building, number of floors in the building, building's height, surface area and XY coordinates in Lambert 93 CRS.

OpenStreetMap (OSM) is a collaborative project that aims at creating free maps of the entire world. We extracted buildings from OSM[43] and considered "housing" all buildings tagged as residential and alike, plus those that didn't have any tag or name to be recognized else how (the majority)[44].

Then, we linked other tags used for the buildings to NA64 codes. NA64 code (Activité économique en 64 postes) can be found in the MOBZELT table and represents 64 different economic activities. Most likely economic activities were recognized and linked to buildings extracted from the OSM database. Tags from OSM were used as well as key words from easily recognizable names, such as "École" (French word for "school"), "Musée" (museum), "Piscine" (pool), etc.

- ***SCHOOLPLACE (Adresse et géolocalisation des établissements d'enseignement du premier et second degrés)***

The dataset of the location of schools and universities is provided by the Ministère de l'Éducation Nationale, de la Jeunesse et des Sports[45]. Based on "codes nature" provided for each school, we paired schools to the corresponding suitable age range.

[43] https://wiki.openstreetmap.org/wiki/Buildings
[44] https://wiki.openstreetmap.org/wiki/Tag:building%3Dapartments
[45] https://www.data.gouv.fr/fr/datasets/adresse-et-geolocalisation-des-etablissements-denseignement-du-premier-et-second-degres-1/

A.3 Method

Because of the large amount and size of data, the method has been designed from the beginning to run the algorithm in parallel. We prototyped on a sample IRIS, and later extended the method to all the IRISes running the calculations in parallel for each IRIS. An exception is given by the metropolitan areas of Paris, Lyon and Marseille, for which we used the ARM code (Arrondissement) in place of the IRIS.
A second exception is presented by the "*communes non-découpées"* (see definitions in section A.2.2). These communes are indicated in the tables with the string "ZZZZZZZZZ" in place of the IRIS code. For these communes, we referred to the code "CANTVILLE", which is the coarser census unit. We randomly assigned individuals of a certain cantville to one or another IRISes that belonged to it.
A third exception is the modality defined by the municipality code (or the municipal district code for Paris, Lyon and Marseille) followed by "XXXX", that corresponds to the IRIS of the municipality with less than 200 inhabitants. We treated these records at the level of the first 5 digit code (DEPCOM).
According to this design, in our workflow, the first 5 digits of the IRIS code are given as input.
A propaedeutic step has been merging the information about work mobility into one single table. To do that, starting from the MOBPRO and MOBZELT tables, we have used the intersection features as keys and operated a union of the remaining keys. The result is a WORK_MOB table encompassing all information regarding work mobility.
Based on the given IRIS code, the following datasets have been subset in order to reduce the size of the input files: INDCVI, LOGEMT, HOUSING, WORK_MOB (previously merged), WORKPLACE. The resulting files, together with SCHOOLPLACE and CONTOURS IRIS, are loaded and fed as input to the model.

A.3.1. Enrichment of people profiles with workplace data

In the INDCVI table, as mentioned earlier, each record contains characteristics of an individual and a weight (IPONDI) that gives a measure of the frequency with which every record is found in the population in a certain census area.
The profile of an individual is characterised by a set of n features:

$$INDCVI = <f_1, f_2, ..., f_n>$$

The WORK_MOB table

$$WORK\_MOB = <f_{wm1}, f_{wm2}, ..., f_{wmM}>$$

contains some features overlapping the INDCVI table, that we used as keys for merging:

$$merging\_keys = INDCVI \cap WORK\_MOB$$

The added attributes are those present in WORK_MOB table that were not present in INDCVI table:

$$added\_attributes = WORK\_MOB \text{ XOR } INDCVI$$

The resulting table INDCVI' is given by:

$$INDCVI' = INDCVI \cup added\_attributes$$

One of the added attributes is DCLT (Département, commune et arrondissement municipal (Paris, Lyon, Marseille) du lieu de travail). Based on this attribute, we could associate the location (census area) of the workplace to the individuals. We then made a step further and considering also the N64 code (Activité économique en 64 postes) randomly picked one workplace from the destination census area that met the N64 code, and associated such workplace to the individual in terms of coordinates.

A.3.2 Adding attributes to students

Similarly to what has been done for work mobility, the MOBSCO table has been intersected to the INDCVI' table:

$$merging\_keys = INDCVI' \cap MOBSCO$$

and the added attributes:

$$added\_attributes = MOBSCO \text{ XOR } INDCVI'$$

merged to INDCVI', generating INDCVI''.

In the process, additional merging keys were also generated, namely:

- AGEREV10 (Age révolu en 10 classes) is an attribute present in MOBSCO but not in INDCVI'. Before merging, we reclassified the AGEREV (Âge en années révolues détaillé) attribute from INDCVI to obtain AGEREV10.
- INEEM (Nombre d'élèves, étudiants ou stagiaires âgés de 14 ans ou plus du ménage) is also an attribute present in MOBSCO but not in INDCVI. However, we know from INDCVI the attributes needed to generate it: ETUD (Inscription dans un établissement d'enseignement - yes / no) and AGEREV (age).
- In MOBSCO, the attribute DCETUF (Département, commune et arrondissement municipal (Paris, Lyon, Marseille) du lieu d'études) represents the census area of the school / university in which the individual is enrolled. Using this attribute, combined with the age, it is possible to shortlist a set of educational institutes from the SCHOOLPLACE table. The selection among them is operated on a random basis, as we could not find information on the actual number of enrolled students per each educational institute.

A.3.3 Individuals unweighting

As explained earlier, since the data come from statistical surveys, all calculations must be carried out with the weight of the individual (IPONDL for LOGEMT, IPONDI for the other tables). In this step, the individuals are disaggregated and families are created, generating a family ID (famid) attribute.

A.3.4 House - family mapping

In this step, we map families, created in the previous step, into houses from the HOUSING file. This combinatorial optimization problem is known as the Variable Size Multiple Knapsack Problem. Some authors (Thiriot & Sevenet, 2020) propose a probabilistic approach to pair households to housing. This problem can be tackled in different ways, no solution is perfect but there is always a trade-off between precision and computational intensity. Aiming at a better precision is only possible when the input data add up useful information. Sometimes least computationally intensive solutions offer reasonable results as well. In our case, having any additional attribute to houses, e.g. year when built, would make people positioning much more precise. Another source of uncertainty is that, in lack of better information, we assumed that larger families would inhabit larger housing surfaces, which is obviously not always the case.

Since we do not have any additional attributes, we opted for simplifications that still give balanced results, deterministic and extremely fast: fitting the family size to bins, bins being the house area. After a first tentative, increasing bins if the families do not fit in the current configuration.

A.3.5 Small IRISes

As mentioned earlier, IRISes are defined small if the area covers less than 200 people. These IRISes are indicated with XXXX in the original data. In these cases, a larger area has been considered, incorporating adjacent census areas.

A.3.6 Enrichment of profiles with POIs

Based on the aforementioned BPA dataset, starting from the locations of home, workplace / education, we assigned the most likely points of interests to individuals. The closest facilities to home, workplace / school were filtered according to the distance and assigned to the record of each individual.

A.3.7 The parallel processing

The algorithm has been applied over the whole French territory, taking as a computational unit the census area (IRIS or arrondissement for the metropolitan areas). The calculation has been performed on the JRC in-house Big Data Analytics Platform BDAP (Soille et al. 2018), suitable for large-scale computations[46], that uses HTCondor as a job scheduler and Docker Universe set up.
The batch processing job has run over 500 nodes, each node running a Docker container equipped with the image created ad-hoc with all the software and libraries needed (essentially scripts in Bash and Python 3, including pandas, geopandas, numpy, scipy, shapely etc.).
Around 35k jobs were run, one for each French commune, each job taking 1 CPU. At our disposal were 20 servers of 40 CPUs each and 1TB RAM, and relatively unlimited storage space. The machine set was shared with other users.

A.4 Quality check

We performed a quality check on the city of Lille (DEPCOM = 59350). First, we compared some general statistics against the official ones[47].

	Synthetic population estimate	Official statistics (2018)
Population Lille	230655	233098
Number of households	135734	121429
Percentage of salaried jobs on the total workers	91%	91.3%

We also compared the synthetic population against the High Resolution Population Density Map from Facebook For Good[48]. This latter, for France refers to census data collected on 2009. This dataset individuated 7 categories: population, males, females, females aged 15 to 49, children aged 0 to 5, young people aged 15 to 24, elderly people aged 60 or older.

In the following figure, on the left is the split according to Facebook For Good dataset, and on the right is the synthetic dataset.

[46] https://jeodpp.jrc.ec.europa.eu/services/shared/home/
[47] https://statistiques-locales.insee.fr/#c=report&chapter=compar&report=r01&selgeo1=com_courant.59350
[48] https://dataforgood.facebook.com/dfg/tools/high-resolution-population-density-maps

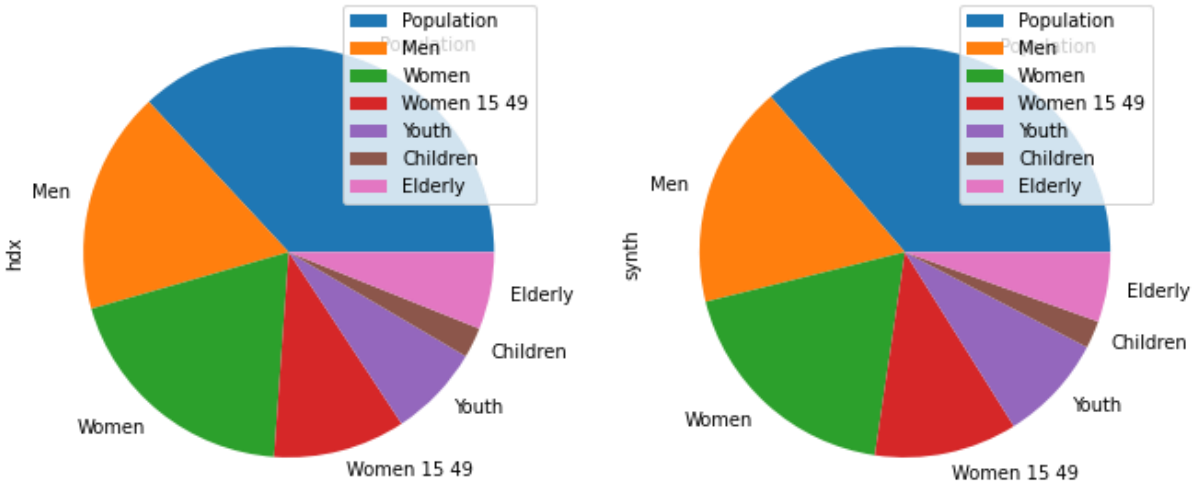


Figure 49 Data split according to Facebook For Good dataset (left) vs. generated synthetic dataset (right).

A.5 Data availability

The dataset has been under scrutiny now and will be released for unlimited use for all Commission's services on BDAP platform.

Annex II. The MOSTLY. AI full report

Full QA Report

generated on 07 Dec 2021, 01:14AM

for 456,919 target vs. 1,500,000 synthetic subjects

MOSTLY·AI

MOSTLY·AI

Executive Summary

Overall Accuracy: 98.8% ✓ Privacy Tests: ✓

Table	Columns	Target Data	Synthetic Data	Accuracy
patient	4 categorical 2 datetime	456,919 rows	1,500,000 rows	univariate: 99.3% bivariate: 98.9%
cancer_split_TMBG	3 numeric 5 categorical	479,883 rows	1,564,450 rows	univariate: 99.3% bivariate: 98.5% autocorrelation: 99.2%

Privacy Tests

Identical Match Share (IMS)

✓

Holdout Data
IMS: 0.14% = 15 of 10000 subjects were found to have identical matches in the target data.

Synthetic Data
IMS: 0.02% = 3 of 10000 subjects were found to have identical matches in the target data.

TEST PASSED! The share of subjects within the synthetic data that matches an actual subject from the target data is not significantly bigger than the share that is to be expected when analyzing the target data itself.

Distance to Closest Record (DCR)

✓

Holdout Data
5^{th} percentile of the DCR = 0.02

Synthetic Data
5^{th} percentile of the DCR = 0.02

TEST PASSED! The normalized distance for synthetic subjects to their closest actual subject within the target data is not significantly closer than the distance that is to be expected when analyzing the target data itself.

Nearest Neighbor Distance Ratio (NNDR)

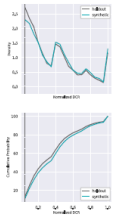
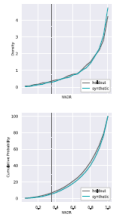
✓

Holdout Data
5^{th} percentile of the NNDR = 0.3

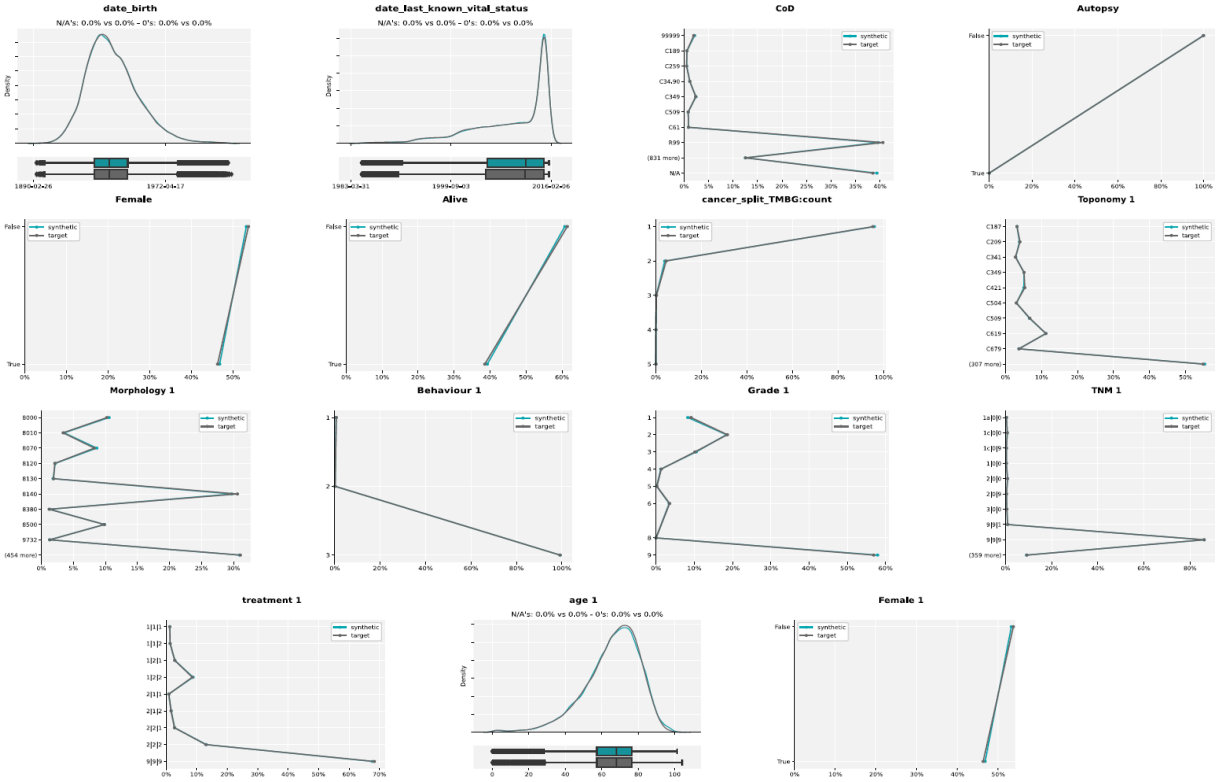
Synthetic Data
5^{th} percentile of the NNDR = 0.34

TEST PASSED! The distance ratio between nearest and second-nearest record for synthetic subjects to their closest actual subject within the target data is not significantly closer than the ratio that is to be expected when analyzing the target data itself.

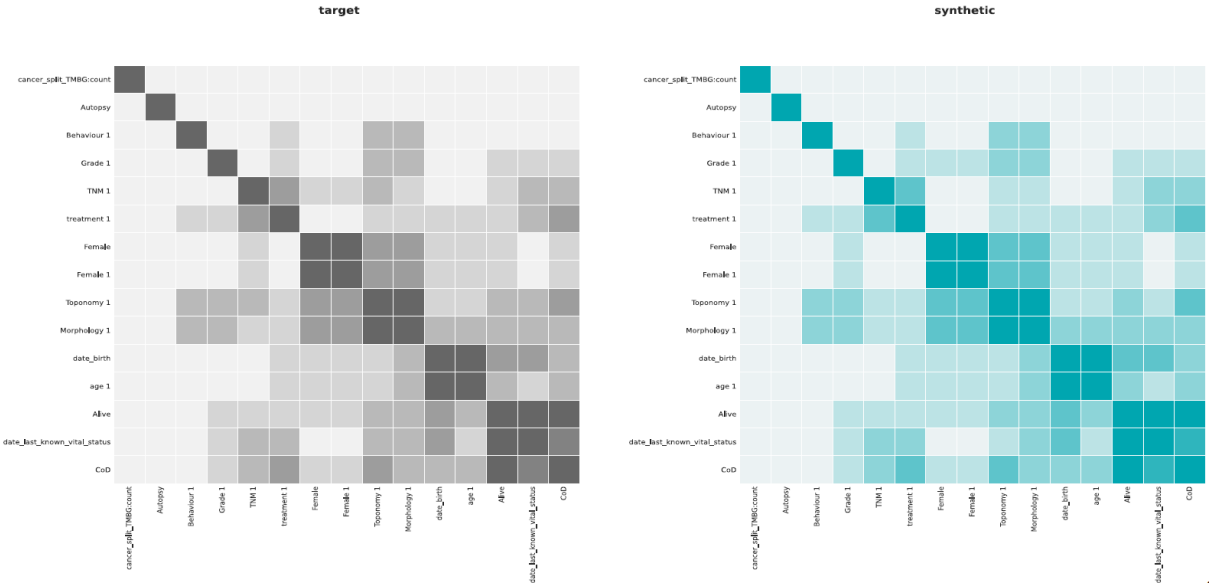
Privacy Tests

Identical Match Share (IMS)	Distance to Closest Record (DCR)	Nearest Neighbor Distance Ratio (NNDR)
✓	✓	✓
Holdout Data: 0.1415% (15 of 10000)		
Synthetic Data: 0.0215% (3 of 10000)		

Univariate Distributions



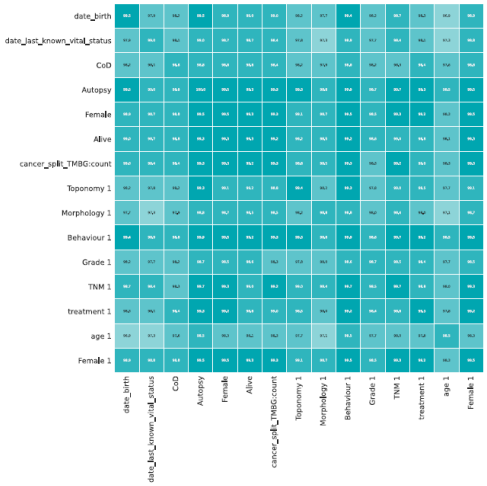
Correlations



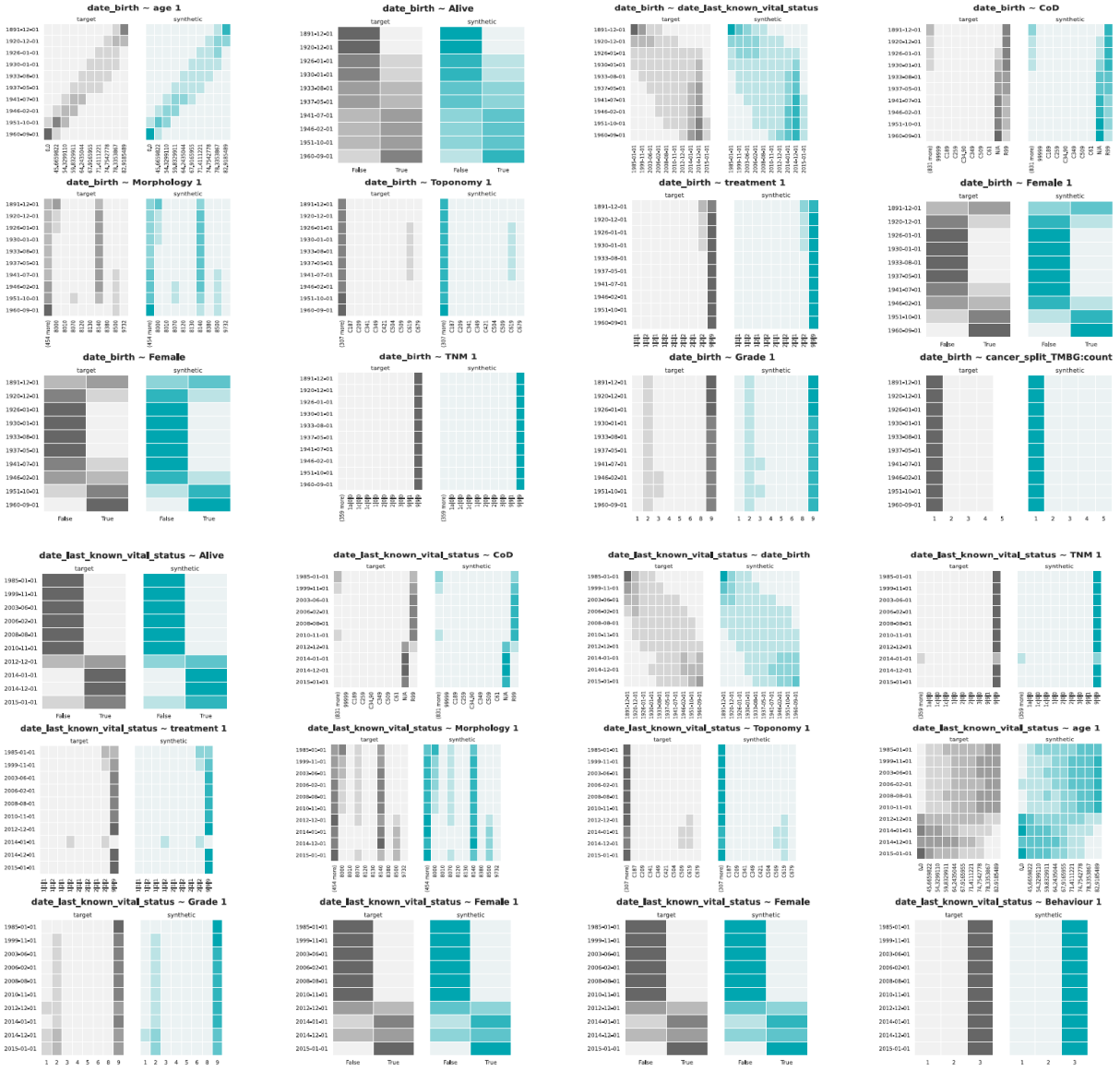
Accuracy

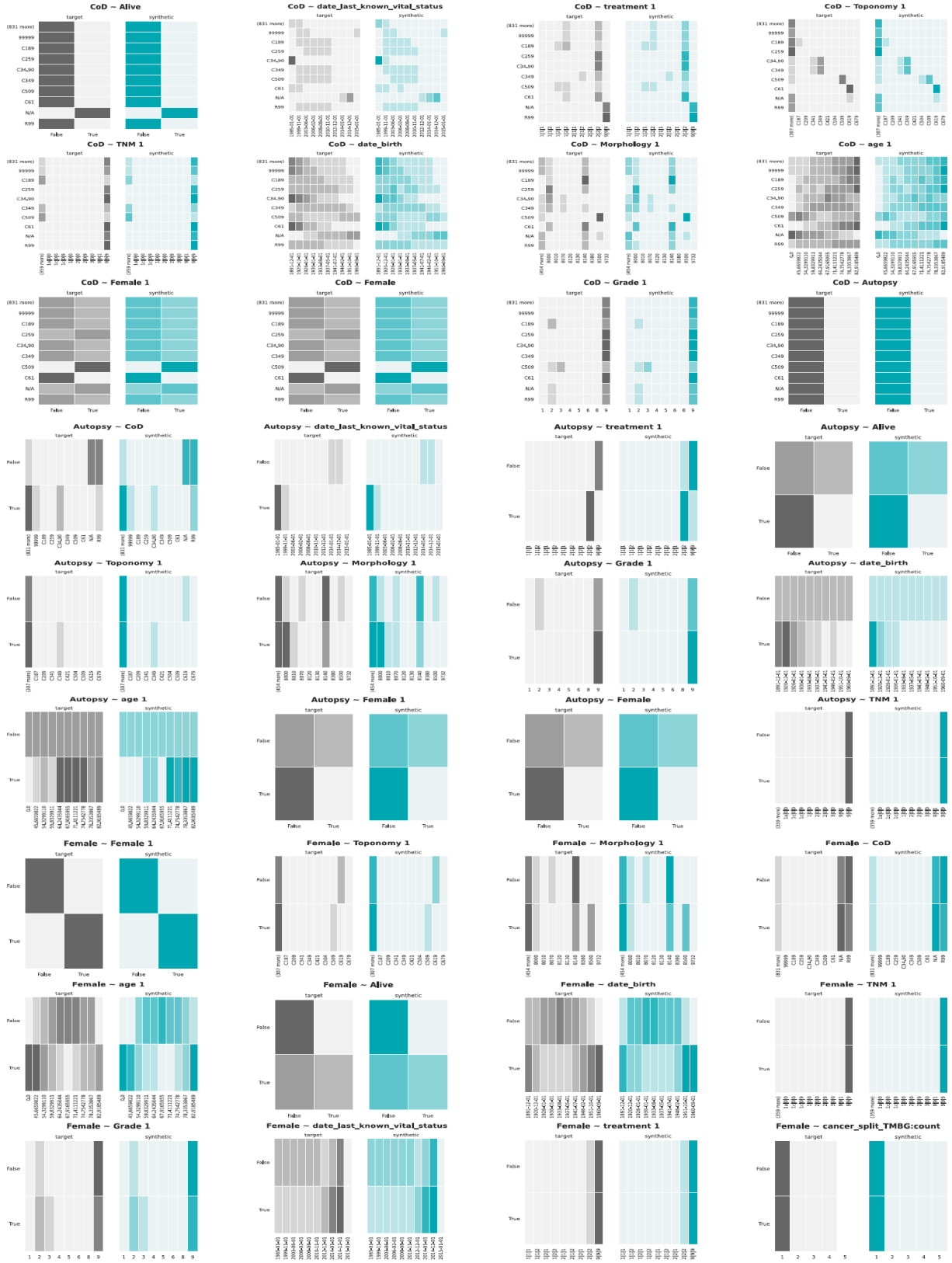
Accuracy 'target ~ synthetic': 98.78% [96.88%, 99.99%]

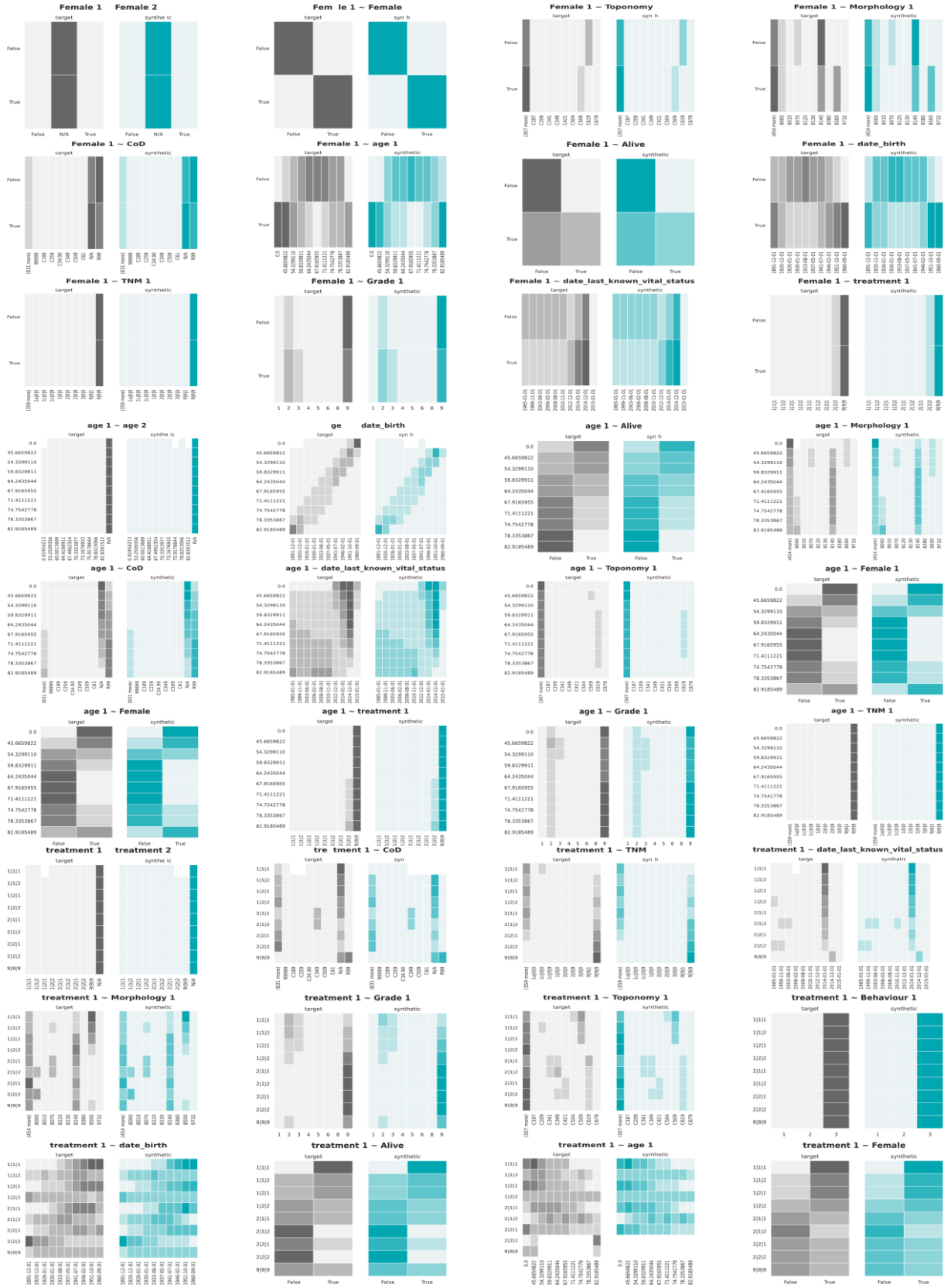
Accuracy is defined as '100% - (Deviation in Probability)'



Bivariate Distributions




CoD ~ Alive
CoD ~ date_last_known_vital_status
CoD ~ treatment 1
CoD ~ Toponomy 1
CoD ~ TNM 1
CoD ~ date_birth
CoD ~ Morphology 1
CoD ~ age 1
CoD ~ Female 1
CoD ~ Female
CoD ~ Grade 1
CoD ~ Autopsy
Autopsy ~ CoD
Autopsy ~ date_last_known_vital_status
Autopsy ~ treatment 1
Autopsy ~ Alive
Autopsy ~ Toponomy 1
Autopsy ~ Morphology 1
Autopsy ~ Grade 1
Autopsy ~ date_birth
Autopsy ~ age 1
Autopsy ~ Female 1
Autopsy ~ Female
Autopsy ~ TNM 1
Female ~ Female 1
Female ~ Toponomy 1
Female ~ Morphology 1
Female ~ CoD
Female ~ age 1
Female ~ Alive
Female ~ date_birth
Female ~ TNM 1
Female ~ Grade 1
Female ~ date_last_known_vital_status
Female ~ treatment 1
Female ~ cancer_split_TMBG:count


Female 1
Female 2
Fem le 1 ~ Female
Female 1 ~ Toponomy
Female 1 ~ Morphology 1
Female 1 ~ CoD
Female 1 ~ age 1
Female 1 ~ Alive
Female 1 ~ date_birth
Female 1 ~ TNM 1
Female 1 ~ Grade 1
Female 1 ~ date_last_known_vital_status
Female 1 ~ treatment 1
age 1 ~ age 2
ge
date_birth
age 1 ~ Alive
age 1 ~ Morphology 1
age 1 ~ CoD
age 1 ~ date_last_known_vital_status
age 1 ~ Toponomy 1
age 1 ~ Female 1
age 1 ~ Female
age 1 ~ treatment 1
age 1 ~ Grade 1
age 1 ~ TNM 1
treatment 1
treatment 2
tre tment 1 ~ CoD
treatment 1 ~ TNM
treatment 1 ~ date_last_known_vital_status
treatment 1 ~ Morphology 1
treatment 1 ~ Grade 1
treatment 1 ~ Toponomy 1
treatment 1 ~ Behaviour 1
treatment 1 ~ date_birth
treatment 1 ~ Alive
treatment 1 ~ age 1
treatment 1 ~ Female

References

Aaberge, R., Bach, S., Bargain, O., Burlacu, I., Buslei, H., Capéau, B., ... & Decoster, A. (2014). Handbook of Microsimulation Modelling.

Alemanno, A., Amir, O., Bovens, L., Burgess, A., Lobel, O., Whyte, K., & Selinger, E. (2012). Nudging healthy lifestyles–informing regulatory governance with behavioural research. European Journal of Risk Regulation, 3(1).

Alemanno, A., & Spina, A. (2014). Nudging legally: On the checks and balances of behavioral regulation. International Journal of Constitutional Law, 12(2), 429-456.

Almagro, M. (2022). Political polarization: Radicalism and immune beliefs. Philosophy & Social Criticism. https://doi.org/10.1177/01914537211066859

Antoni, J. P., Vuidel, G., & Klein, O. (2017). Generating a located synthetic population of individuals, households, and dwellings. Luxembourg Institute of Socio-Economic Research (LISER) Working Paper, (2017-07).

Arendt, H. (1972). Crises of the Republic: Lying in Politics, Civil Disobedience on Violence, Thoughts on Politics, and Revolution. New York: Harcourt Brace Jovanovich. ISBN 978-0-15-623200-5.

Arentze, T., Timmermans, H., & Hofman, F. (2007). Creating synthetic household populations: problems and approach. Transportation Research Record, 2014(1), 85-91.

Beaulieu-Jones, B. K., Wu, Z. S., Williams, C., Lee, R., Bhavnani, S. P., Byrd, J. B., & Greene, C. S. (2019). Privacy-preserving generative deep neural networks support clinical data sharing. Circulation: Cardiovascular Quality and Outcomes, 12(7), e005122.

Beckman, R. J., Baggerly, K. A., & McKay, M. D. (1996). Creating synthetic baseline populations. Transportation Research Part A: Policy and Practice, 30(6), 415-429.

Bellovin, S. M., Dutta, P. K., & Reitinger, N. (2019). Privacy and synthetic datasets. Stan. Tech. L. Rev., 22, 1.

Bindschaedler, V., Shokri, R., & Gunter, C. A. (2017). Plausible Deniability for Privacy-Preserving Data Synthesis (Extended Version). arXiv preprint arXiv:1708.07975.

Bosco, C., Grubanov-Boskovic, S., Iacus, S., Minora, U., Sermi, F. and Spyratos, S., Data Innovation in Demography, Migration and Human Mobility, EUR 30907 EN, Publications Office of the European Union, Luxembourg, 2022, ISBN 978-92-76-44518-0 (online),978-92-76-46702-1 (print), doi:10.2760/027157 (online),10.2760/958409 (print), JRC127369.

Brick, C., Freeman, A. L., Wooding, S., Skylark, W. J., Marteau, T. M., & Spiegelhalter, D. J. (2018). Winners and losers: communicating the potential impacts of policies. Palgrave Communications, 4(1), 1-13.

Burgard, J. P., Kolb, J. P., Merkle, H., & Münnich, R. (2017). Synthetic data for open and reproducible methodological research in social sciences and official statistics. AStA Wirtschafts-und Sozialstatistisches Archiv, 11(3), 233-244.

Charness, G., & Chen, Y. (2020). Social identity, group behavior, and teams. Annual Review of Economics, 12, 691-713.

Chen, S. H., & Pollino, C. A. (2012). Good practice in Bayesian network modelling. Environmental Modelling & Software, 37, 134-145.

Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F., & Sun, J. (2017, November). Generating multi-label discrete patient records using generative adversarial networks. In Machine learning for healthcare conference (pp. 286-305). PMLR.

Ciriolo, E. (2011). Behavioural economics in the European Commission: past, present and future. Oxera Agenda, 1-5.

Citro, C. F., & Hanushek, E. A. (Eds.). (1991). Improving information for social policy decisions: the uses of microsimulation modeling volume 1 review and recommendations. National Academy Press.

Collins, D. (2015). The Oxford Handbook of Behavioral Economics and the Law by Eyal Zamir and Doron Teichman (eds) Oxford: Oxford University Press, 2014, 824 pp.,€ 143, 39; Hardback. European Journal of Risk Regulation, 6(3), 470-472.

Craglia, M., Scholten, H.J., Micheli, M., Hradec, J., Calzada, I., Luitjens, S., Ponti, M. and Boter, J., (2021) Digitranscope: The governance of digitally-transformed society, EUR 30590 EN, Publications Office of the European Union, Luxembourg, ISBN 978-92-76-30229-2, doi:10.2760/503546, JRC123362.

De Groeve, T., Ratto, M., Annunziato, A., Hradec, J., Benczur, P. and Le Blanc, J. (2020) JRC COVID-19 de-escalation modelling: framework for linking health and socio-economic factors, European Commission, 2020, JRC120602

Delhoum, Y., Belaroussi, R., Dupin, F., & Zargayouna, M. (2020). Activity-Based Demand Modeling for a Future Urban District. Sustainability, 12(14), 5821.

Deming, W. E., & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. The Annals of Mathematical Statistics, 11(4), 427-444.

Ding, N., Zarrabian, M. A., & Sadeghi, P. (2021). $\alpha $-Information-theoretic Privacy Watchdog and Optimal Privatization Scheme. arXiv preprint arXiv:2101.10551.

Dingel, J. I., & Neiman, B. (2020). How many jobs can be done at home?. Journal of Public Economics, 189, 104235.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2011). Differential privacy—a primer for the perplexed,". Joint UNECE/Eurostat work session on statistical data confidentiality, 11.

El Emam, K. (2020). Seven ways to evaluate the utility of synthetic data. IEEE Security & Privacy, 18(4), 56-59.

Farooq, B., Bierlaire, M., Hurtubia, R., & Flötteröd, G. (2013). Simulation based population synthesis. Transportation Research Part B: Methodological, 58, 243-263.

Gargiulo, F., Ternes, S., Huet, S., & Deffuant, G. (2010). An iterative approach for generating statistically realistic populations of households. PloS one, 5(1), e8828.

Girin, L., Leglaive, S., Bie, X., Diard, J., Hueber, T., & Alameda-Pineda, X. (2020). Dynamical variational autoencoders: A comprehensive review. arXiv preprint arXiv:2008.12595.

Goodfellow, I. (2016). Nips 2016 tutorial: Generative adversarial networks. arXiv preprint arXiv:1701.00160.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S. & Bengio, Y. (2014). Generative adversarial nets. Advances in neural information processing systems, 27.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

Hitaj, B., Ateniese, G., & Perez-Cruz, F. (2017, October). Deep models under the GAN: information leakage from collaborative deep learning. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (pp. 603-618).

HLP (2013). 'A New Global Partnership: Eradicate poverty and transform economies through sustainable development', The Report of the High-Level Panel of Eminent Persons on the Post-2015 Development Agenda. United Nations Publications.

Hsu, H., Asoodeh, S., & Calmon, F. P. (2019, July). Information-theoretic privacy watchdogs. In 2019 IEEE International Symposium on Information Theory (ISIT) (pp. 552-556). IEEE.

Huang, C., Kairouz, P., Chen, X., Sankar, L., & Rajagopal, R. (2017). Context-aware generative adversarial privacy. Entropy, 19(12), 656.

Ironmonger, D., Jennings, V., & Lloyd-Smith, B. (2000, October). Long term global projections of household numbers and size distributions for LINK countries and regions. In Project LINK meeting, Oslo, Norway on (pp. 3-6).

Jorosz, B. (2013). Estimating Households by Household Size Using the Poisson Distribution. In Population Association of America.

Kagho, G. O., Ilahi, A., Balać, M., & Axhausen, K. W. (2020). Synthetic population of Greater Jakarta: An iterative proportional updating approach. In 20th Swiss Transport Research Conference (STRC 2020) (virtual). STRC.

Kaloskampis, I., Joshi, C., Cheung, C., Pugh, D., & Nolan, L. (2020). Synthetic data in the civil service. Significance, 17(6), 18-23.

Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P., & Sanil, A. P. (2006). A framework for evaluating the utility of data altered to protect confidentiality. The American Statistician, 60(3), 224-232.

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. In International Conference on Learning Representation (ICLR).

Kolb, S. K., Münnich, R., & Templ, M. (2011) Synthetic Data Generation of SILC Data.

Kosti, N., Levi-Faur, D., & Mor, G. (2019). Legislation and regulation: three analytical distinctions. The theory and practice of legislation, 7(3), 169-178.

Lenormand, M., & Deffuant, G. (2012). Generating a synthetic population of individuals in households: Sample-free vs sample-based methods. arXiv preprint arXiv:1208.6403.

Liu, J., Ma, X., Zhu, Y., Li, J., He, Z., & Ye, S. (2021). Generating and Visualizing Spatially Disaggregated Synthetic Population Using a Web-Based Geospatial Service. Sustainability, 13(3), 1587.

Machanavajjhala, A., He, X., & Hay, M. (2017, May). Differential privacy in the wild: A tutorial on current practices & open challenges. In Proceedings of the 2017 ACM International Conference on Management of Data (pp. 1727-1730).

Mair, D., Smillie, L., La Placa, G., Schwendinger, F., Raykovska, M., Pasztor, Z. and Van Bavel, R. (2019) Understanding our Political Nature: How to put knowledge and reason at the heart of political decision-making, EUR 29783 EN, Publications Office of the European Union, Luxembourg, ISBN 978-92-76-08620-8, doi:10.2760/910822, JRC117161.

Nalisnick, E., Matsukawa, A., Teh, Y. W., Gorur, D., & Lakshminarayanan, B. (2019). Do Deep Generative Models Know What They Don't Know?. In International Conference on Learning Representations.

Namazi-Rad, M. R., Mokhtarian, P., & Perez, P. (2014). Generating a dynamic synthetic population–using an age-structured two-sex model for household dynamics. PloS one, 9(4), e94761.

Near, J. P., & Abuah, C. (2021). Programming Differential Privacy. URL: https://uvm-plaid.github.io/programming-dp/

Patki, N., Wedge, R., & Veeramachaneni, K. (2016, October). The synthetic data vault. In 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA) (pp. 399-410). IEEE.

Percival, Donald B.; Walden, Andrew T. (1993). Spectral Analysis for Physical Applications. Cambridge University Press.

Peng, R. D. (2011). Reproducible research in computational science. Science, 334(6060), 1226-1227.

Platzer, M., & Reutterer, T. (2021). Holdout-based empirical assessment of mixed-type synthetic data. Frontiers in big Data, 4.

Raab, G. M., Nowok, B., & Dibben, C. (2021). Assessing, visualizing and improving the utility of synthetic data. arXiv e-prints, arXiv-2109.

Reiter, J. P., Wang, Q., & Zhang, B. (2014). Bayesian estimation of disclosure risks for multiply imputed, synthetic data. Journal of Privacy and Confidentiality, 6(1).

Rezende, D. J., Mohamed, S., & Wierstra, D. (2014, June). Stochastic backpropagation and approximate inference in deep generative models. In International conference on machine learning (pp. 1278-1286). PMLR.

Saadi, I., Mustafa, A., Teller, J., Farooq, B., & Cools, M. (2016). Hidden Markov Model-based population synthesis. Transportation Research Part B: Methodological, 90, 1-21.

Sajjad, M., Singh, K., Paik, E., & Ahn, C. W. (2016). A data-driven approach for agent-based modeling: simulating the dynamics of family formation. Journal of Artificial Societies and Social Simulation, 19(1), 9.

Soille, P., Burger, A., De Marchi, D., Kempeneers, P., Rodriguez, D., Syrris, V., & Vasilev, V. (2018). A versatile data-intensive computing platform for information retrieval from big geospatial data. Future Generation Computer Systems, 81, 30-40.

Squires, G. D., & Kubrin, C. E. (2005). Privileged Places: Race, Uneven Development and the Geography of Opportunity in Urban America. Urban Studies, 42(1), 47–68. http://www.jstor.org/stable/43096213

Stadler, T., Oprisanu, B., & Troncoso, C. (2020). Synthetic Data–Anonymisation Groundhog Day. arXiv preprint arXiv:2011.07018.

Steel, P. (2007). The nature of procrastination: a meta-analytic and theoretical review of quintessential self-regulatory failure. Psychological bulletin, 133(1), 65.

Stevens, F. R., Gaughan, A. E., Linard, C., & Tatem, A. J. (2015). Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. PloS one, 10(2), e0107042.

Sun, L., & Erath, A. (2015). A Bayesian network approach for population synthesis. Transportation Research Part C: Emerging Technologies, 61, 49-62.

Taichman, D. B., Sahni, P., Pinborg, A., Peiperl, L., Laine, C., James, A., ... & Backus, J. (2017). Data sharing statements for clinical trials: a requirement of the International Committee of Medical Journal Editors. Annals of Internal medicine, 167(1), 63-65.

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. European journal of social psychology, 1(2), 149-178.

Tanton, R. (2018). Spatial microsimulation: developments and potential future directions. International Journal of Microsimulation, 11(1), 143-161.

Tappin, B. M., & McKay, R. T. (2019). Investigating the relationship between self-perceived moral superiority and moral behavior using economic games. Social Psychological and Personality Science, 10(2), 135-143.

Thiriot, S., & Sevenet, M. (2020). Pairing for Generation of Synthetic Populations: the Direct Probabilistic Pairing method. arXiv preprint arXiv:2002.03853.

Triastcyn, A., & Faltings, B. (2019). Generating Artificial Data for Private Deep Learning. In Proceedings of the PAL: Privacy-Enhancing Artificial Intelligence and Language Technologies, AAAI Spring Symposium Series (No. CONF).

Voas, D., & Williamson, P. (2001). Evaluating goodness-of-fit measures for synthetic microdata. Geographical and Environmental Modelling, 5(2), 177-200.

Woo, M. J., Reiter, J. P., Oganian, A., & Karr, A. F. (2009). Global measures of data utility for microdata masked for disclosure limitation. Journal of Privacy and Confidentiality, 1(1).

Wood, A., Altman, M., Bembenek, A., Bun, M., Gaboardi, M., Honaker, J., ... & Vadhan, S. (2018). Differential privacy: A primer for a non-technical audience. Vand. J. Ent. & Tech. L., 21, 209.

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional gan. Advances in Neural Information Processing Systems, 32.

Ye, X., Konduri, K., Pendyala, R. M., Sana, B., & Waddell, P. (2009, January). A methodology to match distributions of both household and person attributes in the generation of synthetic populations. In 88th Annual Meeting of the Transportation Research Board, Washington, DC.

List of abbreviations and definitions

AI	Artificial Intelligence
API	Application Program Interface
DT	Digital Twin
IA	Impact Assessment
NLP	Natural Language Processing
PSP	Probabilistic Synthetic Population

List of figures

List of tables

GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: https://europa.eu/european-union/contact_en

On the phone or by email

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),

- at the following standard number: +32 22999696, or

- by electronic mail via: https://europa.eu/european-union/contact_en

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website at: https://europa.eu/european-union/index_en

EU publications
You can download or order free and priced EU publications from EU Bookshop at: https://publications.europa.eu/en/publications. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see https://europa.eu/european-union/contact_en).

KJ-NA-31116-EN-N

The European Commission's science and knowledge service

Joint Research Centre

JRC Mission

As the science and knowledge service of the European Commission, the Joint Research Centre's mission is to support EU policies with independent evidence throughout the whole policy cycle.

EU Science Hub

ec.europa.eu/jrc

@EU_ScienceHub

EU Science Hub - Joint Research Centre

EU Science, Research and Innovation

EU Science Hub

Publications Office of the European Union

doi:10.2760/50072

ISBN 978-92-76-53478-5