Data shortage for urban energy simulations? An empirical survey on data availability and enrichment methods using machine learning?

Gerald Schweiger[a], Johannes Exenberger[a], Avichal Malhotra[b], Thomas Schranz[a], Theresa Boiger[a], Christoph van Treeck[b], James O'Donnell[c]
[a] Graz University of Technology, Austria, [b] RWTH Aachen University, Germany, [c] University College Dublin, Ireland
gerald.schweiger@tugraz.at

Abstract. Building energy simulations at district and urban scales are vital to design and operate sustainable energy systems. In many cases, these simulations rely on enrichment methods as the required detailed data on building characteristics are often unavailable. Approaches using machine learning to address this problem have already been proposed in the literature. However, research on this topic is still at an early stage and the question of whether machine learning can offer substantial solutions has not yet been answered. The goal of this work is twofold; based on an expert survey, we identify the main challenges regarding data availability for urban energy simulations. Furthermore, we identify possibilities of machine learning methods in the field of data enrichment and city information models to offer an initial contribution in defining further research perspectives in this domain.

1. Introduction

The building sector is responsible for around 40% of total final energy consumption in the European Union (European Commission, 2019) and holds enormous potential for saving energy and reducing CO2 emissions in a cost-effective way. In recent years, building energy demand simulations on district or urban scale have become an increasingly relevant topic in academic research and practical applications. Energy performance simulations are crucial for (a) energy management and control, (b) the design of smart systems to reduce overall energy consumption and (c) the design of solutions for efficiently incorporating new sources of renewable energy within the supply system (Schweiger *et al.*, 2020). 3D city models are vital for energy simulations, as they provide information about buildings in a standardized manner. An overview of models and formats can be found in (Hong *et al.*, 2020; Malhotra *et al.*, 2021).

As detailed data about those building characteristics is often not available (especially on district and urban scale) most modelers enrich models with data from other sources (Malhotra *et al.*, 2020). Another way to enrich building-related data is the inference of certain building features from other features using Machine Learning (ML) techniques. In general, data enrichment can be classified into two main categories: a) the enrichment of geometry, and b) the enrichment of semantic data. The first category includes all approaches that use enrichment to create more complex 3D models through data enrichment. ML has for instance been used to identify roof geometries from LiDAR data (Biljecki and Dehbi, 2019). Semantic data enrichment, on the other hand, includes all approaches that identify additional building features that are stored as attributes within the geometrical model. Henn et al. (Henn *et al.*, 2012) for example, use ML for building type classification from a LOD1 city model. Using a different approach, von Platten et al. (Von Platten *et al.*, 2020) combine ML and expert knowledge to identify building types from Google Street View images for estimating energy retrofitting potential.

ML based enrichment methods are a new and emerging field, making it necessary to define potential applications and research paths. As ML cannot be discussed without an assessment of the availability of required data sources, this paper therefore envisions:

- to identify the main challenges researchers are facing regarding data availability for urban energy simulations.
- to identify potential applications for ML in enriching data for district and urban energy simulations.

2. Method

An exploratory expert survey was conducted to explore data availability and enrichment methods using ML for urban energy simulations. Expert surveys are usually conducted in cases where experts have knowledge that is not yet available in the scientific community and the public (Flick *et al.*, 2018). The empirical methodology is similar to the one in (Skov *et al.*, 2021) and (Schweiger, Kuttin and Posch, 2019). We selected academic experts based on (i) their number of publications on city information modeling that are listed in the literature database Scopus and (ii) their active involvement in international projects on city information modeling. Practitioners were chosen according to their actual involvement in projects using city information modeling. 44 experts received the link to an online survey constructed with the survey tool Lime Survey (Limesurvey, 2021),leading to a total number of 28 complete answers. Thus, the response rate was 64%. The questionnaire consists of 18 questions ranging from simple yes/no questions to Likert-scale questions and short-answer questions. To accommodate for additional answers, an extra open field was provided where appropriate. The results of the quantitative questions are presented in a bar chart and, if applicable, evaluated in terms of median and mean, which ensures a transparent presentation of the results.

There are clear limitations associated with the method that was applied in this paper. A well-known problem in interviewing experts is the representativeness of the sample population (Christopoulos, 2011). Exploratory expert surveys gather facts and information to explore new research topics or to establish an initial orientation in a nascent field (Flick *et al.*, 2018). In general, the method implies rather small sample sizes. Since exploratory expert surveys do not aim at generalization, there is no special requirement to have a representative sample or even to interview all relevant experts (Kaiser, 2014). Helfferich, for example, recommends interviewing between 6 and 30 experts (Helfferich, 2011).

3. Results

The first question of the survey concerns the field of applications of city information models for researchers and practitioners working in the domain of urban energy simulation (see Figure 1). The majority of the respondents (70%) have been or are currently using city information models for heating demand prediction of buildings, with an additional 19% of respondents planning to do so within the next year. The second major application for city information models is the visualization of energy demand, which was already done by 65% of the respondents and planned by another 31%. More than half of all respondents have applied digital city models in the context of electric energy (58%) and cooling energy (52%) demand prediction and simulation. Although currently not used as frequent as applications for heating demand prediction, 27% and 30% plan to use city information models for electricity and cooling demand computations respectively. This trend correlates with the increasing importance of cooling systems for overall energy consumption in the future due to rising temperatures, especially in urban areas. Optimal planning and operation of energy production were not considered as applications for city information models by as many respondents. 35% use city models for optimal planning, 38% intend to do so in the coming 12 months. For optimal

operation, 38% worked or are working with such models, while 23% are planning to work on this topic in combination with city information models. The data from the survey does not allow statements about the general importance of individual research topics, but the results indicate that the use of city information models is less beneficial for optimal planning and control, at least given the current state of the art methods.

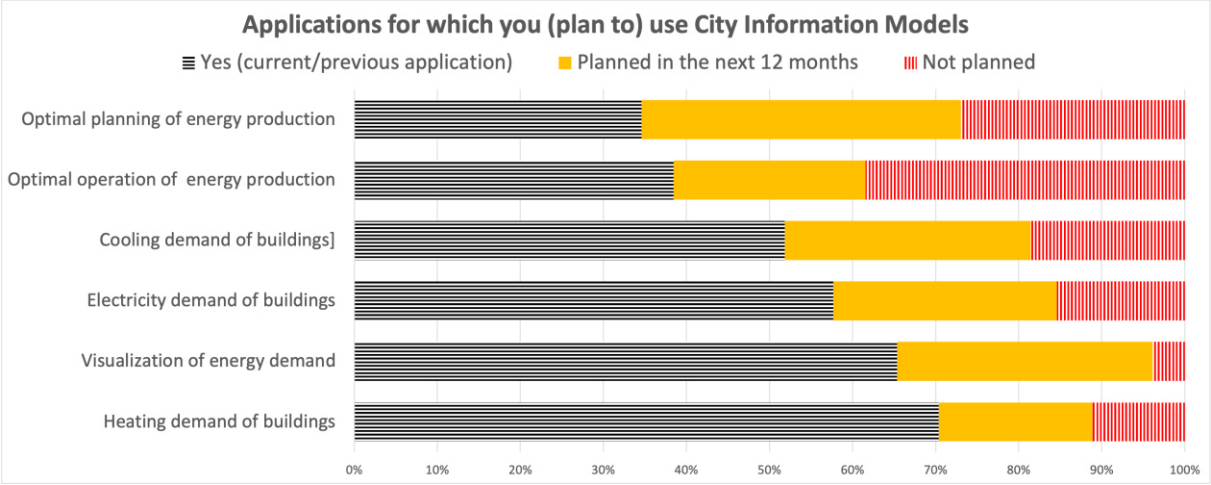

Figure 1: Field of applications for city information models

About one third of the experts mentioned additional applications that were not included in the survey. The applications mentioned can be categorized into the following main objectives: reduction of greenhouse gas emissions; research on urban heat islands; agent-based modeling and traffic modeling; simulation of energy networks; other simulations (noise, pollution, ...); urban planning and smart cities.

In the following question (Figure 2), the respondents are asked how time and effort in a their projects is usually distributed across the following work packages: data acquisition, development of the simulation model, simulation and results analysis. This was done to identify existing bottlenecks in the workflow of projects regarding urban energy simulations, highlighting potential applications for ML. Data acquisition is considered the most time consuming part of the workflow, with an average of 44% (median = 40%) of the whole project time dedicated to data acquisition. Additionally, more than 40% of all respondents spend at least half of their time on data acquisition, indicating that acquiring and pre-processing data is still a major bottleneck regarding research in the domain of urban energy simulations. The development of a simulation model accounts for less time according to the respondents (average = 30%, median = 25 %). Performing the simulation and analyzing the results on average makes up for 27% of the overall time consumption, with a median value of 37%, indicating that across all respondents, variations in time consumption are largest for this work package.

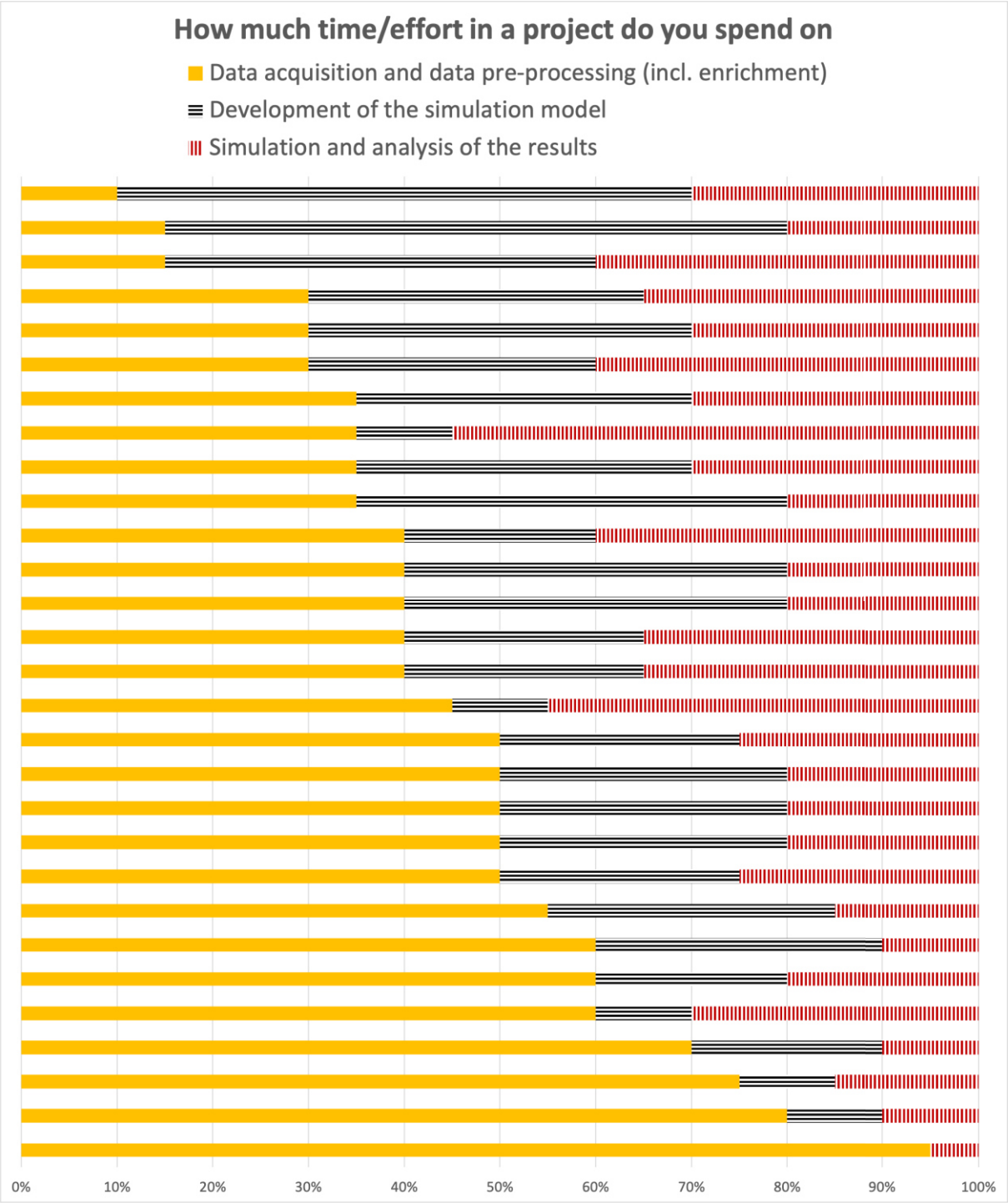

Figure 2: Distribution of workload across different phases of a project

For energy performance applications, efficient use of technology has also proven to be an important milestone in the development of workflows offering sustainable energy management solutions. Methods such as ML or image processing have already been used for building energy assessments. However, these are generally limited to individual buildings and lack implementations on an urban scale. Furthermore, for urban energy simulations, efficient usage of virtual 3D city models along with the previously mentioned methods can be a big step towards energy efficient districts. Virtual data models at a city scale are generally limited in their availability. As the landscape of available data sources is quite complex, with varying

restrictions for usage and publication, an approach by Malhotra, et al. (Malhotra *et al.*, 2020) categorizes different availability types of data for energy-related applications. Using these categories, the participants of the present survey were asked to name the types of data they are frequently using (see Figure 3).

All the participants agreed to use open source datasets, whereas, only 18% acknowledged the use of commercial data sets. Commercial data refers to information that is licensed and can be used by paying an agreed fee. Moreover, 89% of the respondents utilize public sector information where a charge may apply for a certain usage. Academic data that is free of charge for scientific research studies is used by 68% of the participants. Industry restricted data that can only be used for specific applications is used by 29%. Furthermore, 68% of the respondents do not acknowledge the usage of private data that is not available to the people outside an institution, university or industry. Conclusively, as a majority of participants rely on open data sets and public sector information, it is quite important for governmental organizations to make urban scale data publicly available for urban energy applications.

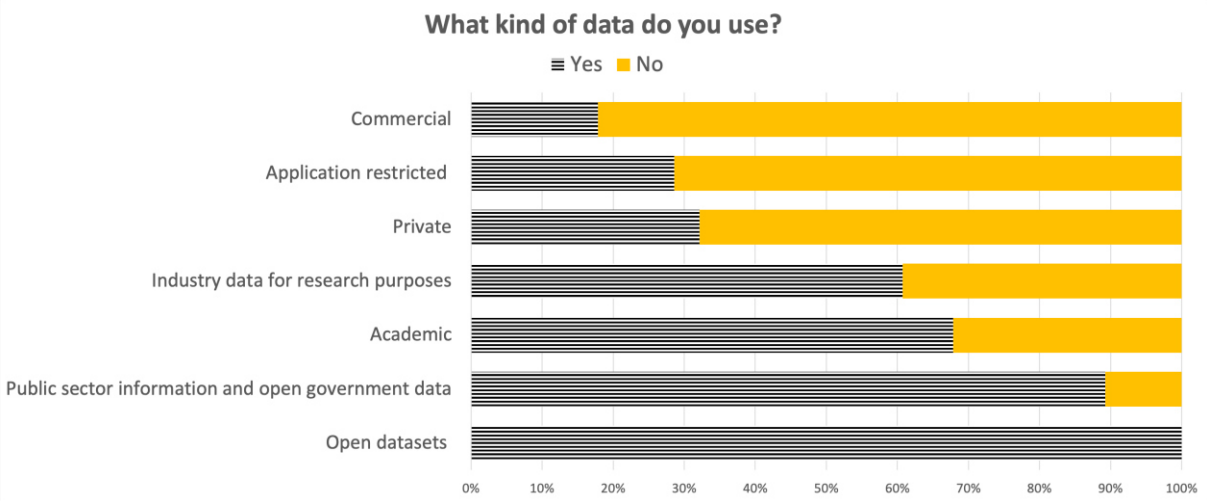

Figure 3: Types of data sources used

Geometrical and energy-specific data is the core requirement for energy related applications. Though many different data models and formats exist, some of them are prominently used in the field of urban building energy modeling (UBEM). City Geographical Markup Language (CityGML) (Gröger *et al.*, 2012), an open XML-based data format, facilitates the representation of semantical and topological information in 3D city models. Although some cities and municipalities offer open LoD1-2 CityGML datasets, there still exists a lack of data models for many different urban areas. Furthermore, CityGML datasets mainly contain geometrical information of the buildings. To include additional information, these models can also be extended using the Application Domain Extension (ADE) mechanism. For energy-relevant information, the CityGML Energy ADE (Agugiaro *et al.*, 2018) is mainly used. Green Building XML (gbXML) (Cheng and Das, 2014), an open data format, also supports information exchange between BIM models and other related analysis tools. Furthermore, the Industry foundation Classes (IFC) (Laakso and Kiviniemi, 2012) can also be used for representing 3D BIM models. The GeoJSON (Dorman, 2020), based on JavaScript Object Notation, defines JSON objects and their relation by which they are combined to represent data about geographic features, their properties and their spatial extents. The ESRI Shapefile format is a geospatial vector data format for geographic information system (GIS) software (ESRI, 1998).

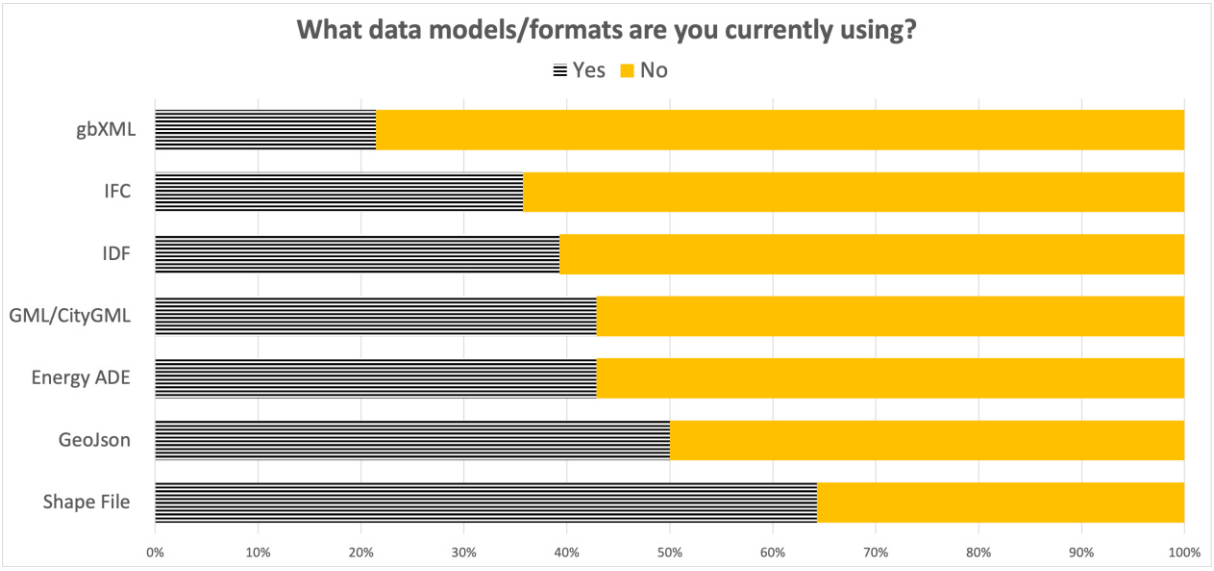

Figure 4: Data models/formats used by the respondents

Although many other data models and formats exist for energy related applications, the ones used most prominently were considered in the survey (see Figure 4). Half of the respondents acknowledged the usage of GeoJSON, whereas the CityGML and Energy ADE are utilized by 43% of the participants. 64% also agreed to use shape files for urban scale applications. Furthermore, IFC and Input Data File (IDF) were selected by 36% and 39% of the respondents respectively. Only 21% considered the usage of gbXML. Furthermore, some experts mentioned the usage of csv, xml, GeoPackage files (gpkg), ESRI File Geodatabase (GDB), Digital elevation models, Geotiff, OpenDRIVE, the UtilityNetwork ADE, 3D pointclouds, 3D meshes, glTF, COLLADA, KML and 3DTiles.

The next section of the questionnaire concerns the topic of ML and data enrichment. Results from the survey show that 82% of experts use data enrichment methods (see Figure 5). Archetype approaches are applied by 75% of the study participants, statistical approaches by 50% and ML methods by 36%. The high percentage of participants using ML methods for data enrichment is surprising, given the relatively low number of publications concerning this topic. Besides these approaches, the participants mentioned other enrichment methods such as engineering models, expert guessing and manual enrichment.

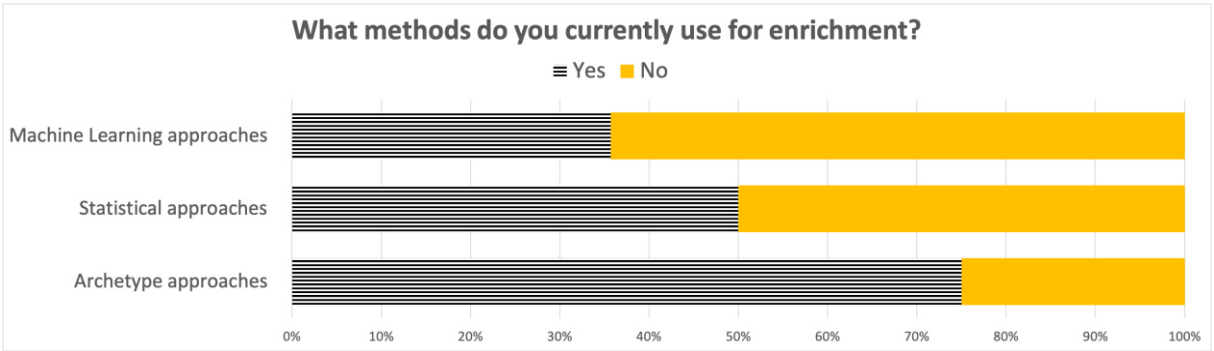

Figure 5: Data enrichment methods used by the participants

68% of experts answered that they have already applied ML techniques in their work. From the remaining 32% percent who have not yet used ML methods, 78% said that they plan to do so

in the future. With a share of 86%, Python is the language/framework of choice for most experts. Matlab and R are used by 21% and 25% respectively. Other languages, such as C++ were only mentioned once. ML can be used in a variety of tasks, such as data pre-processing, data analysis or enrichment. 39% of experts use ML methods for pre-processing, 36% use it for input data analysis and data enrichment and 32% use it to analyze simulation results. Other applications mentioned only by one participant each are modeling and LiDAR image processing. Most experts identify a moderate to high potential for ML techniques in all of these areas (see Figure 6). In data enrichment and input data analysis ML is considered to have a high potential by 64% and 61% of the survey participants respectively. Moderate potential in data enrichment is identified by 28% of the experts and 30% consider ML to have moderate potential in input data analysis. 52% of experts see high potential for ML in data pre-processing and within the simulation workflow (e.g. in the form of surrogate modeling). Moderate potential in these two areas is identified by 32% and 29% respectively. In post-processing and in the analysis of the simulation results 42% of the experts see high potential for ML methods and another 42% see moderate potential.

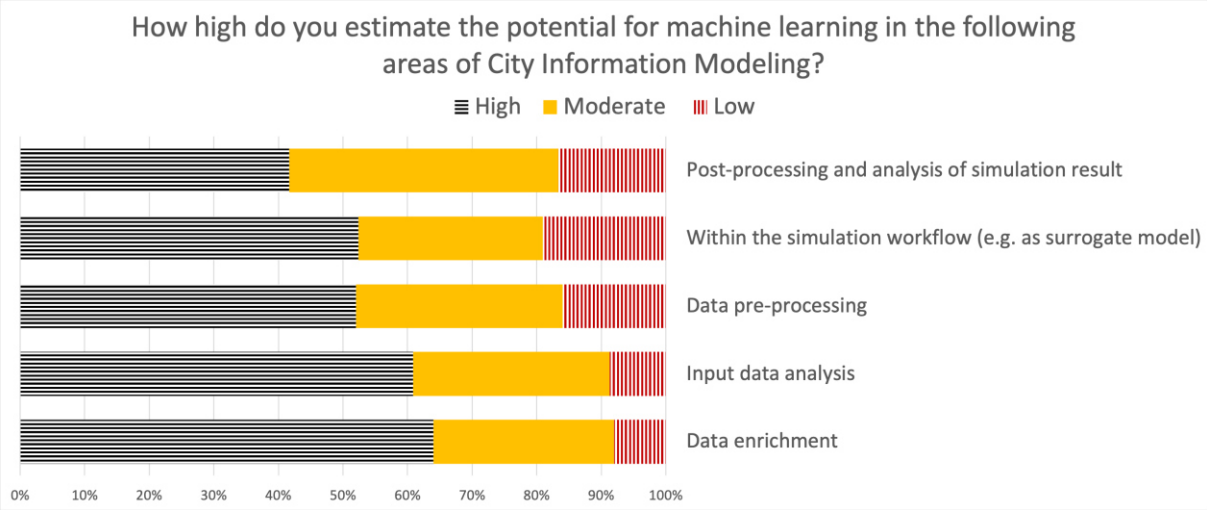

Figure 6: Estimation of potential areas for ML in the domain of city information modeling.

In a following question, the respondents were asked about specific applications in the context of data enrichment they consider promising for the integration of ML. From all the answers, three main topics can be derived: parameter estimation and filling of data gaps, creation of more precise archetypes and image analysis. A majority of experts see the potential of ML in tackling the problems of missing or fragmented data. Closely related is the creation of more accurate building archetypes from data that subsequently can be used to enrich city information models. Image analysis was mentioned several times as well, although exact use cases for image analysis were not specified in most answers. Two experts mentioned image recognition in context of textures for city information models and the detection of building attributes such as windows and PV systems. Data calibration and quality checking was also considered by some respondents. Interestingly, the use of ML for occupancy estimation was only mentioned by one respondent in the survey.

When asked about the potential of ML in the domain of city information models in general, the answers show a less clear opinion across the respondents. While many experts acknowledge the potential of ML for a variety of applications for city information models, many do not settle on definitive use cases, indicating that research in this domain is still in its early stages. Data analysis and processing was also mentioned by several experts. The potential of ML for

generative tasks was also considered, proposing the use of ML for 3D model reconstruction from point clouds and meshes and the generation of imaginary 3D data for planning purposes. Using ML in combination with city information models for energy demand prediction was also mentioned by several respondents.

4. Conclusion

District and city energy simulations are vital to design and operate sustainable energy systems. This paper presents an expert assessment on data availability and potentials for ML techniques to enrich data. The main findings from this paper are:

- Data acquisition is considered the most time-consuming part of the workflow, with a median of 40% of the whole project time dedicated to this task.
- All experts who participated in the survey use open-source datasets, whereas only 18% acknowledged the use of commercial data sets.
- More than 80% of experts use data enrichment methods; archetype approaches are applied by 75%, statistical approaches by 50% and ML methods by 36%.
- Most experts identify a medium to high potential of ML techniques for pre- and post-processing, in the simulation workflow, and for input data analysis and data enrichment. Experts expect the highest potential for ML to be in the area of data enrichment and input data analysis. Three main topics can be derived regarding specific applications in the context of data enrichment: parameter estimation and filling of data gaps, creation of more precise archetypes and image analysis.

It can be concluded that many experts consider ML a promising approach for data enrichment in the domain of urban energy simulation. The number of respondents already using ML for this purpose was higher than the authors expected, given the relatively few publications in this field. On the other hand, fragmented data and the complete lack of available sources still persist as a significant limiting factors for researchers. This is also reflected in the survey, with many respondents having to dedicate the biggest share of their time available for a project to data acquisition. This situation puts the use of ML in a different perspective, as data availability is a crucial requirement for the development of functioning ML approaches. While ML thus has potential for many applications in the domain of city information modeling and urban energy simulation, solving the problem of an absence of useful data cannot be addressed merely through ML.

Acknowledgement

This work emerged from the IBPSA Project 1(Wetter *et al.*, 2019), an international project conducted under the umbrella of the International Building Performance Simulation Association (IBPSA). Project 1 will develop and demonstrate a BIM/GIS and Modelica Framework for building and community energy system design and operation. The reported research has been conducted within the project KityVR (879419), which has received funding in the framework of ”Stadt der Zukunft”.

References

Agugiaro, G. et al. (2018) ‘The Energy Application Domain Extension for CityGML: enhancing interoperability for urban energy simulations’, Open Geospatial Data, Software and Standards, 5.

Biljecki, F. and Dehbi, Y. (2019) ‘Raise the roof: Towards generating LOD2 models without aerial surveys using machine learning’, in ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences. doi: 10.5194/isprs-annals-IV-4-W8-27-2019.

Cheng, J. C. P. and Das, M. (2014) ‘A bim-based web service framework for green building energy simulation and code checking’, Journal of Information Technology in Construction.

Christopoulos, D. C. (2011) ‘Towards Representative Expert Surveys: Legitimizing the Collection of Expert Data’, SSRN Electronic Journal. doi: 10.2139/ssrn.1353283.

Dorman, M. (2020) ‘GeoJSON’, in Introduction to Web Mapping. doi: 10.1201/9780429352874-7.

ESRI (1998) ‘ESRI Shapefile Technical Description’, Computational Statistics. doi: 10.1016/0167-9473(93)90138-J.

European Commission (2019) Energy performance of buildings. Available at: https://ec.europa.eu/energy/en/topics/energy-efficiency/energy-performance-of-buildings.

Flick, U. et al. (2018) ‘Generating Qualitative Data with Experts and Elites’, in The SAGE Handbook of Qualitative Data Collection. doi: 10.4135/9781526416070.n41.

Gröger, G. et al. (2012) ‘OGC City Geography Markup Language (CityGML) En-coding Standard’, Ogc.

Helfferich, C. (2011) Die Qualität qualitativer Daten, Die Qualität qualitativer Daten. doi: 10.1007/978-3-531-92076-4.

Henn, A. et al. (2012) ‘Automatic classification of building types in 3D city models’, GeoInformatica. doi: 10.1007/s10707-011-0131-x.

Hong, T. et al. (2020) ‘Ten questions on urban building energy modeling’, Building and Environment. doi: 10.1016/j.buildenv.2019.106508.

Kaiser, R. (2014) Qualitative Experten-interviews: Konzeptionelle Grundlagen und praktische Durchf hrung, Springer.

Laakso, M. and Kiviniemi, A. (2012) ‘The IFC standard - A review of history, development, and standardization’, Electronic Journal of Information Technology in Construction.

Limesurvey (2021) LimeSurvey: An Open Source survey tool.

Malhotra, A. et al. (2020) ‘A review on country specific data availability and acquisition techniques for city quarter information modelling for building energy analysis’, in BauSIM 2020.

Malhotra, A. et al. (2021) ‘City Quarter Information Modeling for Building Energy - ATaxonomic Review’, Under Review.

Von Platten, J. et al. (2020) ‘Using machine learning to enrich building databases-methods for tailored energy retrofits’, Energies. doi: 10.3390/en13102574.

Schweiger, G. et al. (2020) ‘Active consumer participation in smart energy systems’, Energy & Buildings.

Schweiger, G., Kuttin, F. and Posch, A. (2019) ‘District heating systems: An analysis of strengths, weaknesses, opportunities, and threats of the 4GDH’, Energies, 12(24). doi: 10.3390/en12244748.

Skov, I. R. et al. (2021) ‘Power-to-X in Denmark: An Analysis of Strengths, Weaknesses, Opportunities and Threats’, Energies. Multidisciplinary Digital Publishing Institute, 14(4), p. 913.

Wetter, M. et al. (2019) ‘IBPSA Project 1: BIM/GIS and Modelica framework for building and community energy system design and operation - Ongoing developments, lessons learned and challenges’, in IOP Conference Series: Earth and Environmental Science. doi: 10.1088/1755-1315/323/1/012114.