International Journal of Sustainable Energy Planning and Management

RESEARCH and EXPERIMENTATION

Supporting tool for multi-scale energy planning through procedures of data enrichment

Francisco Javier Miguel-Herrero*, Víctor Iván Serna-González and Gema Hernández-Moral

Fundación CARTIF, Parque Tecnológico de Boecillo, 205, 47151 Boecillo, Valladolid, Spain

ABSTRACT

Considering the challenge of evaluation of the urban environment from the energy point of view, there is plenty of room to improve the resources currently managed by users, enterprises and public institutions. The goal is to create a tool that supports in the decision making in the energy planning process in specific areas by automatically estimating the energy demand and consumption of buildings using public data and representing the results in a geo-referenced way. The tool will provide a better understanding of what the current status of the buildings is, providing these stakeholders with a larger quantity of useful data about the city environment, including not only the geometric information present in cadastre repositories, but also the data collected from the Energy Performance Certificates (EPCs).

In this case, the data from the cadastre repository are combined with the EPCs for each province, with data about the demanded and consumed energy. The objective is to generate a set of buildings typologies for each province with estimated values for the demand and consumption for each building type. These typologies could be used to generate a map with the energetic values for any municipality of this province.

These results can be injected into GIS (Geographic Information Systems) tools that could show these data in order to evaluate the energy demand/consumption of the municipality easing the energy planning decision-making process, or even into databases for further uses.

1. Introduction

Current worldwide problems such as climate change and growing CO_2 emissions, and the corresponding temperature increase, define a reality where it is necessary to act upon. Fossil fuel consumption needs to be reduced, alternative renewable energy sources should be implemented, as well as energy efficiency strategies promoted and refurbishment of the existing building stock aimed for, since it is one of the most problematic sectors in terms of CO_2 emissions.

All of these strategies need to be implemented at different scales, ranging from the European level (e.g. through the implementation of energy directives), to national and regional scale, and to urban scale, where specific actions can be defined and implemented.

However, these energy planning processes at urban scale are complex, time-consuming and often do not count on the necessary tools to support them, which leads to inadequate assessments and to not progressing at the pace required by the challenges faced.

On the other hand, there is an increasing amount of publicly available data that has not been exploited or put to use to this purpose. In this line, the cadastre and Energy Performance Certificates (EPCs from now on) databases (national and regional respectively) offer a vast amount of data that can result in valuable information to support energy planning processes.

*Corresponding author - e-mail: framig@cartif.es

Thus, the aim of this paper is to present a modular software tool that could cover the foretold goals, adding value to the information contained inside the cadastre repository by means of combining these data with the data obtained in the EPCs, in other words, the energy consumption values and CO_2 emissions contained in the EPCs with the geometry of the buildings included in the cadastre registers, starting at City/Municipality level.

Links to previous works

There are several projects that deal with the mapping of energy demand and consumption for energy planning purposes. Some of them estimate energy consumption at the block and lot level of a determined city [1]. Others focus on the calculation of not only energy consumption, but also energy demand based on calculation methods proposed in Energy Performance Certification processes [2,3], whereas there are other examples which combine this with information coming directly from the Energy Performance Certificates database [4].

However, when dealing with estimations of a large number of building blocks this process is highly time-intensive and resource-consuming; on the contrary, the tool proposed offers estimations that could serve to derive similar conclusions to the abovementioned tools, based on reliable data and with a lower resource cost.

2. Methodology

It is important to highlight that the tool has two clearly differentiated parts. The first part corresponds to the algorithm building in charge of the creation of the building typologies using the information of the public available EPCs in a province. The second part is the application of these typologies in one specific location (municipality) in order to estimate the values for the demand, consumption, primary energy consumption and CO_2 emissions for each building of this location.

For the purposes of the part of the typologies generation, two main data sources were used. On the one hand,

Energy Performance Certificates data from the *Ente Regional de la Energía de Castilla y León* (EREN) through the general open data service from the *Junta de Castilla y León* [5] were acquired, and on the other hand, the other set of data (building data, address, etc.) to be combined came from the Spanish Cadastre [6].

The module's objective is to obtain a combination of both data sources and obtain useful statistics about the energy indicators based on the establishment of typologies based on the use of the building and other parameters such as the climate zone and the year of construction; and also location references for these registers in order to manipulate the data into GIS applications and services.

One of the main reasons to get these data is to try to minimize the effect of the suppositions usually made by other works that rely strongly on simulations or aggregated values [7] (e.g., the aggregations based on the age of the buildings that would not take into account reforms and refurbishment procedures). Thanks to these values, the individual results will be concretely located and the real features of the buildings will be consequently assembled into GIS tools with data derived from EPCs values and not the result of calculated estimations that could not be necessarily accurate.

For the second part of the tool, the estimation of the demanded and consumed energy in one location, the sources of the module are the Spanish cadastre and data about Land Use of Spain (*SIOSE, Sistema de Ocupación del Suelo de España*) [8] a part of the information of the typologies generated. The data from the cadastre are used not only for doing the calculations but also for establishing the location of the building.

The information collected from SIOSE is used in order to complement the information of the current use obtained from the Cadastre.

It is critical to notice that the whole sets of data are publicly available and there are no privacy issues. Otherwise, the procedures should include aggregation procedures to anonymize, like used in [9]

As a summary, all these processes, considering the validation levels for EPCs indicated in [10] will perform validations of levels 0, 1 and 5.

2.1. Typologies generation tool

The main workflow of the tool for the generation of the typologies can be seen in the following figure. The six main phases indicated in the figure are described below.

Phase 1: data setup.
During the initial phase, the goal is to acquire an environment of data files that could help to configure the requests to the cadastre, keeping in mind that these requests represent the largest bottleneck and the most sensitive part of the process.

To aid in our goal of a proper configuration of the requests, some data are necessary:
- **PROVINCES (code list)**. Each and every province of Spain has a distinctive code that has to be gathered in order to make a proper consultation in the cadastre website. The list itself is requested to the cadastre, and a JSON file [11] with the data is recovered and put into the root folder of the application, as long as it will belong to all the consultations. The decision to work with JSON format was taken because a) the format is considered as a valid standard of data and b) the way it is constructed mimics the behaviour of the objects in object-oriented programming that facilitates handling them in languages such as the JavaScript being used.
- **MUNICIPALITIES (code list).** For a single selected province (excluded Basque Country and Navarre because they have their own cadastre system), a data request is performed to get the code values for each municipality in the province. The format of the data is JSON, and amongst other values, the cadastre code number is obtained. The list is put inside a dedicated folder for the province.
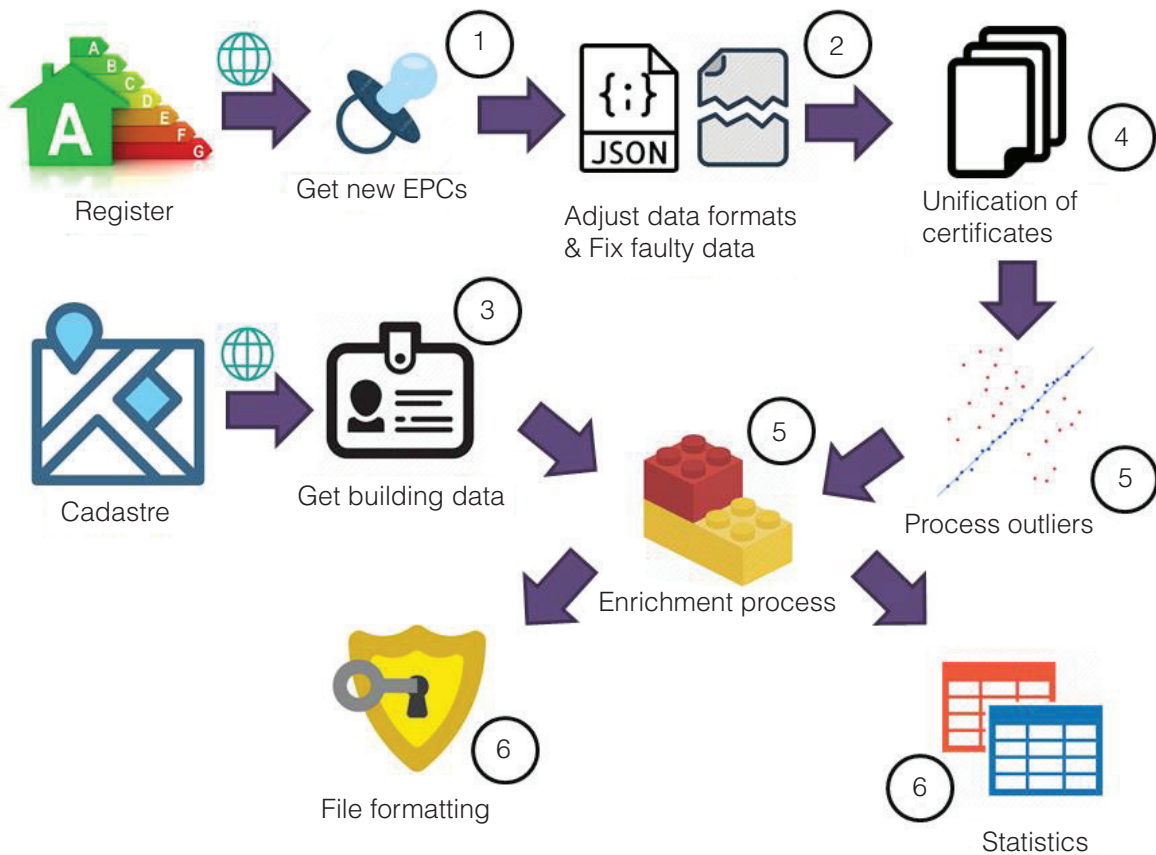

Figure 1: Working flow of the tool developed for the data process

- **STREETS (code list).** For each municipality, a file with the streets codes is downloaded from the cadastre. Everything is inserted into a folder organized like a tree so the files can be easily found.

Phase 2: energy performance certificates acquisition and processing.

The block of data related to energy performance certificates (EPCs) can be also requested via web to a service of open data provided by the Junta de Castilla y León. In Spain, EPC registers are managed at regional level; thus different approaches and data availability may vary from region to region. In other regions there could be similar services that provide the lists with the energy performance certifications necessary for this work, so the module that manages the connection should be adapted.

The system distinguishes between the first instalment and the following ones. First, all the existing certificates are obtained. For the following instalments, only new EPCs need to be processed. However, the system will obtain the full set, so it will have to filter these new ones, thanks to the identification code that the certificates bear.

The next step is to generate the list of objects that will contain the interesting data from the certificates. Most of the variables have a relation one-to-one from the certificates to the destination objects, but some of them need some processing in order to be of use afterwards, concretely, the addresses of the dwellings. The system implemented takes into consideration the habitual layout of the addresses inside the certificates (type of street, name of the street, number, stair/letter/others and postal code in the end) and tries to cope with certain cases and exceptions (suffix in the names of the roads, problems with the codifications, former urban entities absorbed by nearby larger municipalities, etc.). The architecture of this process is purely rule-based, although an alternative considering one customized machine learning tool is proposed as a future development, for example through the usage of TensorFlow [12] or similar software. For that case, the previous experience with the current system will be invaluable to get an initial set of training and evaluation data.

Along with the process of the certification data, the codes referred to the street, the municipality and province are inserted from the code lists obtained during the Phase 1. If there are no coincidences, error fields are filled with the corresponding explanation.

Phase 3: Cadastre identifiers.

The next field to be completed is the cadastre identifier (or Inspire ID) from the element whose energy performance certificate is referred. The quantity of items in use during the latest tests was as large as 30,000 EPCs entries for the province of Valladolid. The system sends one request per object and the cadastre online answers affirmatively or gives off an error message. During this process, the objects containing the information are split into three different groups: (1) elements processed correctly and that now have their corresponding cadastre id, (2) elements containing some kind of error that have been rejected, and (3) a disappeared category containing some cases that never had a proper answer from the cadastre server. The procedure was performed this way in order to minimize the effect of eventual connection failures to the cadastre, considering that online requests are usually one of the weakest parts from a given process, where there is little to no control for the answers, as long as they are strongly asynchronous and prone to have a large range of different failures. Moreover, the cadastre site in Spain has a protection system against Distributed Denial of Service (DDoS) attacks [13] that forced to limit the number of request per hour in order to avoid being banned/blocked. The key to avoid these problems was to allow some asynchronous behaviour with the connections and the processing of the data request by the means of splitting the code into synchronous/asynchronous processes so the results could be properly ordered, filtered and evaluated, generating checkpoints that could be easily followed for educational, clarity and debugging purposes as well as enabling portability of code, replicability and rearrangement in the order of certain procedures. The last part of this phase is a procedure to reprocess elements that did not get a response from the cadastre server, in order to have the definitive list of elements that can be further processed or not, looking to avoid data holes and incorrect results. It works the same as the general procedure.

Phase 4: The unification of certificates.

Unification of certificates in order to harmonise the input data was necessary. This is the case of existing dwelling certificates inside a building block, which are not comparable to the results that would have been obtained when considering the whole building. In order to reduce the potential discrepancies found among certificates of different dwellings inside a building, in the

tool's approach there is only a single certification element per building representing the mean of all of them for every parameter (demand, consumption, etc.).

One advantage of this unification is the reduction in the number of certification elements that will also reduce the number of web requests in the next phase.

Phase 5: Data enrichment and outlier selection.
Once again, the cadastre is consulted in order to obtain information about the buildings, using the cadastre identifiers that the certificates already have. In the same way that has been commented before, this is another execution bottleneck, very time-consuming and consequently has to be carefully monitored. For this purpose, a modified version of a module created for a previous project [2] has been used.

In this point the system has data from two different sources: the Energy Performance Certificates (containing information about the use of the building and energy data: demand, consumption and CO_2 emissions) and the cadastre data (surface, number of dwellings, year of construction and location for establishing the climate zone). All the data is combined into a single element that contains the important information in terms of consumption, and it is also well referred with the address, coordinates and identifiers.

One important procedure during this phase is the handling of outliers [14] [15]. The outliers can be treated during this part of the process, although the corresponding module is well prepared to work during previous stages of the process as well. The method for eliminating outliers includes the following steps:

- **Cluster generation**: Separation of values into "use of building" clusters or building typologies as they are used in the Spanish cadastre. The categories taking into account the uses of buildings are the following: complete blocks of dwellings; individual homes in building blocks; detached houses; educational facility; commercial building; administrative facility; health and hospitals; sports facilities; hotels and residences; office buildings; and other tertiary usages. In the case of the periods the classification used by the energy performance certification tool CE3X [16] is used, which correspond to relevant changes in building construction regulation: before 1981, from 1981 to 2007, from 2008 to 2012, from 2013 to 2018 and after 2018. For climate-related data, the National Code for Building Construction

[17] in Spain was queried, since it establishes reference climate zones. In our case, for each province only two or three climates zones will be differentiated.

- **Treating small groups of elements**: the small groups of elements have been discarded, since it could make no sense to search for outliers when the number of elements is small. For this case, the number of 50 has been chosen, but it can be changed in-code.
- **Mean and the standard deviation calculation for each set of values**. The values of $\overline{X}$ and σ are obtained, and for every single element of the set (X_i) the following equation is used:

$$\left| \frac{X_{i\text{-}\overline{X}}}{\sigma} \right| > 2.5 \tag{1}$$

The equation works perfectly for the values considered (energy heating and cooling demand) in the current building cases. The values that satisfy the Equation 1 are considered outliers, and the whole element is separated from the general set of values. As it happened to other discarded elements, there is a variable dedicated to indicate the kind of error in order to follow and evaluate these cases.

Phase 6: Data visualization: graphs, tables and GIS-based
From this point on, the system has available some tools that:

- Generate values of aggregated demand, consumption and CO_2 emissions values from every building typology considered.
- Create files containing the data from the energy objects. The input are these objects in JSON format, and the output is a .csv file that would fit perfectly into a table in a database or can be manipulated with an Excel type application.

The Figure 2 and Table 1 show an example of aggregation obtained with the results extracted from the province of Valladolid (Spain).

2.2. Applying the typologies: estimation tool
In a given location not all buildings have an energy certificate, so if we mapped the data directly from EPCs we would only get values for a few buildings.

The generation of the typologies not only allows to obtain a set of typologies that can help to study the
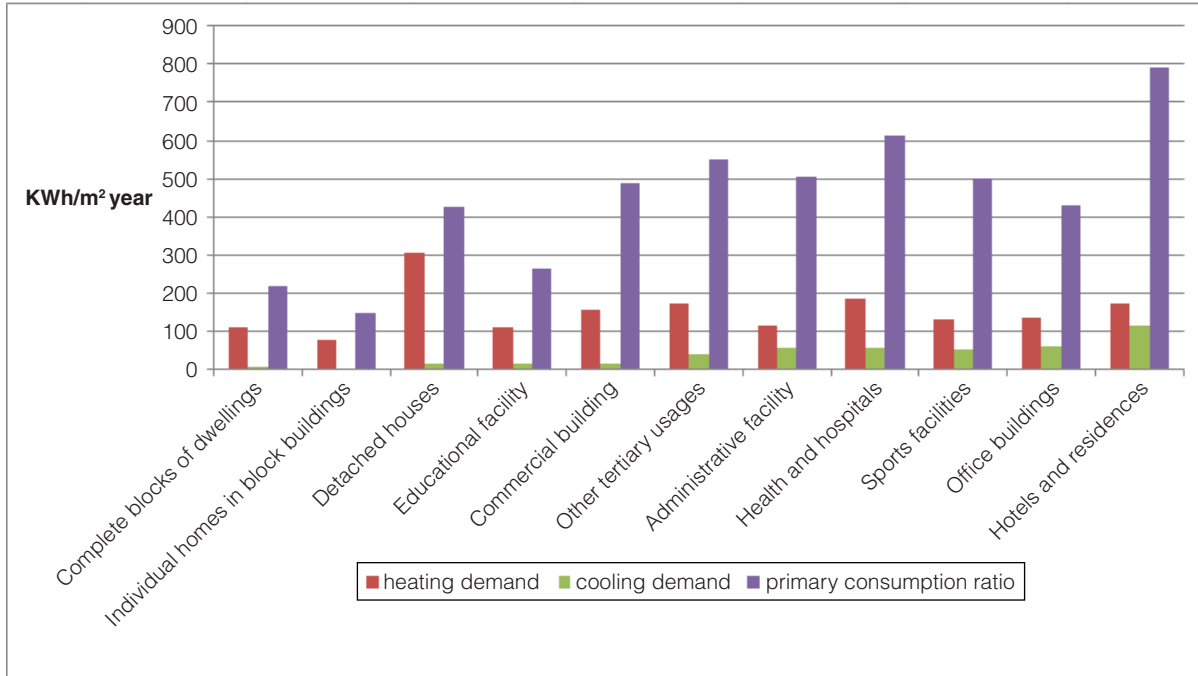
Figure 2: Aggregated values for heating demand, cooling demand and primary energy consumption in kWh/m² per year

Table 1: Example of table of aggregated values for CO_2 emissions per building typology, climate zone and construction period

PERIOD	Climate Zone	Individual dwellings	Residential buildings	Educational facility	Administrative buildings	Health and hospitals	Office buildings	Hotels and residences
before 1981	Y	160.92	40.64	60.60	109.50	134.88	92.25	117,56
1981-2007	Y	82.56	22.50	54.57	121.93	137.62	74,63	283,71
2008-2013	Y	41.34	10.45	88.00	57.00	186.00	122.00	192.00
2014-2018	D	41.14	20.13	0.00	86.00	0.00	0.00	42.00
2014-2018	E	32.80	0.00	0.00	0.00	0.00	0.00	0.00

behaviour of the buildings globally in a province, but also to apply this typologies in one specific location in order to estimate the demand and consumption energy and the emission of the CO_2 of the buildings in this location. The idea is to use the results of the aforementioned process, i.e., the data of demand, consumption and CO_2 aggregated by typology (use, period of construction and climate zone), and to apply them for each building of this location taking into account the use, the year of construction and the climate zone of the building. So an estimation for each building of this location is available regardless of whether or not there is available EPC for that specific building.

Therefore, the estimation based in the application of the typologies on all the buildings can be applied to a determined municipality in order to determine how that city behaves energetically.

For the application of the typologies we need to know different data from the buildings of the location in order to categorize the building in the correspondent typology and also additional information for the calculations. For this purpose data form the Spanish Cadastre and information about land use from SIOSE is used. So the information extracted from each one is:

1. Spanish Cadastre:
   a. Location of the building (for locating in a map and for set the climate zone)
   b. Year of construction
   c. Condition of the building (in order to discard ruined and declined buildings)

Figure 3: Estimation of the CO_2 emissions in the municipality of Medina del Campo using the tool

   d. Surface of the building
   e. Number of dwellings
   f. Use of the building
2. SIOSE:
   a. Current use of the building (in order to complement the information of the Spanish Cadastre because in the case of the building for the tertiary sector SIOSE has information disaggregated: hotels, office, administrative buildings, educational, etc.)

With this information and with the typologies aforementioned this component of the tool is able to create a map of energy demand, consumption and CO_2 emission in GeoJSON format that can be seen in any GIS tool as QGIS [18].

In Figure 3 it can be seen the result of the tool for the Municipality of Medina del Campo (Valladolid) using in the typologies created for the Valladolid province. The values are in tonnes of CO_2 per year. It is important to highlight that the ruin or declined buildings are considered as zero emissions and besides industrial buildings are not categorised.

3. Layout and working tools

The environment selected in order to deal with the tasks proposed was designed with a high degree of flexibility so the design itself would not condition the goals of the overall project. The usage of a programming language like JavaScript enables the use of a simple, well developed and full of libraries language, and Node.js was chosen in order to take advantage of a potent set of built-in modules to simplify the inner programming processes, avoiding spending work hours in existing procedures.

Moreover, the JavaScript fits perfectly when dealing with online transactions, and envisioned future implementations, the developed modules could be easily inserted into websites for user interfaces and open platforms.

The last advantage is the portability into Linux/Unix systems, as long as the current production environment is Windows-based, so the solutions would be considered as compatible with as many platforms as possible.

The code was reworked also to avoid the utilization of non-canonical node libraries. When it was possible, the most generic library was put into use. To name some of them, the *fs* library was utilised for file management, the *cron* library for synchronous-timed calls, http library for Internet connections or *xlsx* library to generate files compatible with Excel. All these libraries are also public and free of charge.

Most of the testing procedures have been performed with average PCs in order to properly evaluate the performance of the algorithms, especially those that would include data combined with large quanti-

ties of requests that have to be checked one by one, keeping in mind that some external issues would appear and cannot be controlled (e.g., internet connection failures).

4. Assessment of outputs and results

The outputs from the whole process (maps and numerical results) have several applications. The most visual result would be the maps. By interacting with them, a user can appreciate what area of a city is most affected in terms of energy demand, energy consumption, primary energy consumption or CO_2 emissions. This estimation and mapping would not be possible had it not been for the typology analysis performed, as well as the identification of these building typologies using two different data sources. Additionally, the numerical results complement this first output by offering more insight on how the building stock has evolved within a certain typology and enables to aim refurbishment strategies to

a specific set of buildings which are more in need than others. In Table 2 results for some typologies in Valladolid are offered, without discriminating per climate zone. In this case, the reduction in time of different values can be observed in most of the cases. Discrepancies in these values can be used as red flags to highlight typologies which need a more in-depth analysis. It must be stated that the main values to be analysed should be the heating and cooling demand, since energy conversion factors affect the results of primary energy consumption and CO_2 emissions. This can potentially lead to misunderstandings when not being able to relate and compare the results to the fuel used by the energy system in a determined building, since this information is not provided as open data.

In addition to the abovementioned applications, the results obtained can be also used in quality checks performed by the authorities in charge of the Energy Performance Certificates, where a value that deviates from the mean of a determined typology or period may

Table 2: Estimation of values per construction period and building typology

Residential buildings	Before 1981	1981–2007	2008–2013	2013–2018	Health and hospitals	Before 1981	1981–2007	2008-2013	2013–2018
Heating demand	98.47	53.81	28.82	60.69	Heating demand	224.29	174.49	80.94	
Cooling demand	4.77	3.77	3.01	8.82	Cooling demand	31.23	63.68	123.31	
Primary Consumption	189.85	105.55	53.63	102.56	Primary Consumption	596.31	594.35	791.00	
CO_2 emissions	41.13	22.77	11.21	21.39	CO_2 emissions	134.88	137.62	186.00	
Individual dwellings	**Before 1981**	**1981–2007**	**2008–2013**	**2013–2018**	**Office buildings**	**Before 1981**	**1981–2007**	**2008–2013**	**2013–2018**
Heating demand	371.64	343.91	108.44	133.45	Heating demand	106.23	124.61	195.39	
Cooling demand	11.90	16.76	8.82	15.10	Cooling demand	53.63	37.81	90.95	
Primary Consumption	691.61	368.34	188.37	199.90	Primary Consumption	399.17	314.13	586.50	
CO_2 emissions	160.92	82.56	41.34	41.14	CO_2 emissions	92.25	74.63	122.00	
Educational facility	**Before 1981**	**1981–2007**	**2008–2013**	**2013–2018**	**Hotels and residences**	**Before 1981**	**1981–2007**	**2008–2013**	**2013–2018**
Heating demand	127.25	89.42	112.19		Heating demand	255.53	91.46	138.72	62.28
Cooling demand	13.31	14.19	44.71		Cooling demand	29.39	213.68	66.76	219.2
Primary Consumption	286.14	234.64	384.00		Primary Consumption	516.89	1219.71	838.00	246
CO_2 emissions	60.60	54.57	88.00		CO_2 emissions	117.56	283.71	192.00	42

imply that the energy performance certificate might need to be revised.

Future evolutions of the tool would be utilized in solutions with larger scope (region-country-world) in order to aggregate energy-related data from various sources (local and global) and the integration with current GIS software (through the usage of GeoJSON or cityGML files, for example).

5. Conclusions

The paper has presented the development and demonstration of a software tool that generates aggregated data from Energy Certificates and Cadastre values, and has some outputs that will be very useful to integrate in GIS energy analysis processes at urban level, as well as for data analysis based on the typology generation.

Acknowledgements

References

[1] Estimated Total Annual Building Energy Consumption at the Block and Lot Level for NYC: http://qsel.columbia.edu/nycenergy/ [retrieved: Nov, 2018]

[2] Hernández Moral, G., et.al. ENERGIS: Tool For Demand Characterisation In Urban Settings To Support Energy Planning At Different Scales. 54th ISOCARP Congress 2018.

[3] Energy Services Platform ENERSI (Plataforma de servicios energéticos basados en la integración y análisis de datos de múltiples fuentes): http://enersi.es/ [retrieved: Nov, 2018]

[4] Energie Label Atlas (Netherlands): http://energielabelatlas.nl/# [retrieved: April, 2018]

[5] Datos Abiertos de Castilla y León. https://datosabiertos.jcyl.es/web/es/datos-abiertos-castilla-leon.html

[6] Sede Electrónica del Catastro. https://www.sedecatastro.gob.es/

[7] X. Oregi et al., Automatised and georeferenced energy assessment of an Antwerp district based on cadastral data. Energy and Buildings, Volume 173, 2018, Pages 176-194, ISSN 0378-7788, https://www.sciencedirect.com/science/article/pii/S0378778817340318

[8] Sistema de Información de Ocupación del Suelo de España. https://www.siose.es/ECMA. Standard ECMA-404. The JSON Data Interchange Syntax. Second Edition. December (2017).

[9] Dochev I, Seller H, Peters I. Spatial aggregation and visualisation of urban heat demand using graph theory. Int J Sustain Energy Plan Manag 2019;24. http://doi.org/10.5278/ijsepm.3346.

[10] Pasichnyi, Oleksii & Wallin, Jörgen & Levihn, Fabian & Shahrokni, Hossein & Kordas, Olga, 2019. "Energy performance certificates — New opportunities for data-enabled urban energy policy instruments?," Energy Policy, Elsevier, vol. 127(C), pages 486-499. https://ideas.repec.org/a/eee/enepol/v127y2019icp486-499.html

[11] ECMA. Standard ECMA-404. The JSON Data Interchange Syntax. Second Edition. December (2017). https://www.ecma-international.org/publications/standards/Ecma-404.htm

[12] Open-source machine learning platform by Google. https://www.tensorflow.org/

[13] Mirkovic J., Reiher P., "A taxonomy of DDoS attack and DDoS defense mechanisms." ACM SIGCOMM Computer Communication Review. Volume 34 Issue 2, April 2004 P. 39–53. https://link.springer.com/chapter/10.1007/978-3-642-11207-2_17

[14] Aguinis, Herman & Gottfredson, Ryan & Joo, Harry. (2013). Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. Organizational Research Methods. 16. 270-–301. 10.1177/1094428112470848. https://journals.sagepub.com/doi/abs/10.1177/1094428112470848?journalCode=orma

[15] Osborne J.W. and Overbay A. The power of outliers (and why researchers should always check for them). North Carolina State University. Practical Assessment, Research and Evaluation. Volume 9, Number 6, March, 2004. https://pareonline.net/getvn.asp?v=9&n=6

[16] Sanctioned document for Energy Certification of Existing Buildings. Developed by CENER & EFINOVATIC. https://www.efinova.es/CE3X

[17] Spanish National Building Code, on Energy Savings (CTE) DB HE: Ahorro de la Energía: https://www.codigotecnico.org/images/stories/pdf/ahorroEnergia/DBHE.pdf [retrieved: Nov 2018]

[18] Free and Open Source Geographic Information System. https://qgis.org/en/site/

[19] Østergaard PA, Maestosi PC. Tools, technologies and systems integration for the Smart and Sustainable Cities to come. Int J Sustain Energy Plan Manag 2019;24. http://doi.org/10.5278/ijsepm.3450