Data-driven statistical analysis of energy performance and energy saving potential in the Flemish public building sector

Yixiao Ma[1,2] **and Glenn Reynders**[1,2]

[1] EnergyVille, Thor Park 8310, 3600 Genk, Belgium
[2] Flemish Institute for Technological Research (VITO), Unit Smart Energy & Built Environment, 2400 Mol, Belgium

E-mail: yixiao.ma@energyville.be

Abstract. This paper will focus on the energy performance of the public building sector in Flanders (Belgium) by analysing the recently published EPC database. The main aim of this paper is to firstly have a qualitive and quantitative overview of the current energy performance of different building categories in the Flemish public building sector. The non-residential building types (office, educational, healthcare etc.) will be categorized. Within each category, a set of typologically representative non-residential buildings will be further identified, while a data driven cluster analysis will be carried out in order to define these non-residential archetypes. Moreover, the less energy efficient building sets will be identified by benchmarking the energy performance within the defined building categories, and the relative potential energy saving targets and pathways will be quantified and evaluated in order to provide policy support in improving energy efficiency of the poorly performing public buildings.

1. Introduction

Publicly owned or occupied buildings represent about 10 – 12% of the total surface area of the EU building stock [1]. In Belgium, non-residential buildings account for 32.5% of the total building stock and comprise a more complex and heterogeneous sector compared to the residential one. The tertiary buildings (public, commercial) account for 27% in total of primary energy consumption of the entire Belgian building stock [2]. In both the Energy Performance of Buildings Directive (EPBD) and the Energy Efficiency Directive (EED), public buildings are addressed to play a critical example-setting role for the new built and refurbished buildings in the Member States [3]. A vast energy saving potential is assumed to be untapped in the public building sector and need be quantified. Energy inefficient public buildings should be identified and prioritized in achieving the renovation targets in cities and municipalities. To estimate the energy savings potential of building stocks, archetypes are often used to represent the whole building stock in bottom-up energy simulation model. Many studies have been focused on developing these archetypes by combining statistical numbers and clustering techniques. A database of harmonized residential building typologies is firstly developed in TABULA project [4]. Descriptive statistics, regression and clustering analysis are more often used to develop residential archetypes [5] [6] [7]. The scarcity of open building-level data sources of the non-residential sector, however, brings an extra level of complexity in conducting similar exercise and having a complete overview of the energy performance of specific types of buildings in this sector. This paper describes the statistical analysis and clustering analysis carried out to identify non-residential building typologies based on the Energy Performance Certificate (EPC) public building database. This database has been released recently [8] and contains more than 9000 public buildings with detailed building energy performance related characteristic data including building construction year, building type, geometrical properties, measured final energy consumption and further geographical information of the public buildings in Flanders, which allows to conduct an in-depth statistical analysis to have an overview on the current energy performance, define archetypes and further calculate energy saving potential. It is however noted that the public building database only contains a fraction of the non-residential building sector. Therefore, this study primarily focuses on administrative office, educational, healthcare, sports, cultural events and public services buildings.

2. Methodology

The methodology consists of three main parts. Firstly, a descriptive statistics analysis is carried out to have an overall understanding of the main variables in the database. Secondly, clustering is conducted by following a typical data science approach: data collection and pre-processing, clustering, and quality check on the cluster results. Lastly, by benchmarking the energy performance, the energy saving potential is calculated accordingly.

2.1. Clustering analysis

2.1.1. Data pre-processing. Building type, building age, useful floor area and measured final energy consumption are selected as key parameters for the cluster analysis. Building age is calculated based on the current year and construction year. Specifically, unlike the theoretically calculated energy performance for residential buildings, energy consumption (heating fuel and electricity) of the public buildings is measured over the period of one year, which adds value of including it in the archetype for calibration purpose. Faulty duplicates, unrealistic extremes values and missing data points are primarily checked manually and removed. Boxplots are generated to understand the distribution of each selected key parameter. Thereafter, the outliers are identified by using the Tukey's fences [9] and removed for the cluster analysis. The data for each the key parameters are standardized before the clustering. Standardization is the process of converting different parameters into the same scale, which further allows to compare values between different types of parameters. Two standardization methods, z-score and min-max, are compared in this study. The z-score is calculated for each original data (x) by comparing with the mean (μ) and standard deviation (σ):

$$z = \frac{x - \mu}{\sigma} \qquad (1)$$

The min-max method normalizes each original data with the minimum and maximum in the dataset:

$$mm = \frac{x - min}{max - min} \qquad (2)$$

2.1.2. Clustering algorithm and performance evaluation. The *k*-means method is one of the most commonly used unsupervised clustering methods for unlabelled data [10]. The *k*-means method first randomly initializes *k* cluster centroids, and each point is assigned to the cluster with the closest centroid to the point, while the centroid is recalculated in each cluster and the loop is repeated until the assignments no longer change clusters between two consecutive iterations. The Elbow method is used for determining the optimal number of clusters by looking at the total intra cluster distance as a function of the number of clusters. The total intra cluster distance is defined as the sum of the distances between the centroid and all points in the cluster. Silhouette method is often used to further evaluate the performance of the cluster analysis and the quality of the clustering results, in the case that the ground-truth labels of the dataset are unknown. The Silhouette Coefficient (*SC*) is defined for each sample data and is composed of the average distance between each sample data and all other points in the same cluster (*a*), and the average distance between each sample data and all other points in the next nearest cluster (*b*). The SC has its range of [-1, 1]. A larger SC always indicates a better defined clustering result [11].

$$SC = \frac{b-a}{max(a,b)} \qquad (3)$$

2.2. Relative energy saving potential

In order to benchmark the energy performance, identify the less energy efficient building sets, the relative energy saving potential is calculated as the difference between the building's current specific final energy consumption ($E_{current}$) and the statistical targeted energy performance ($E_{benchmarking}$) of its corresponding building type, then normalized by current energy consumption. The unit of $E_{current}$ and $E_{benchmarking}$ is kWh/m2/year.

$$Relative\ energy\ saving\ potential = \frac{E_{current} - E_{benchmarking}}{E_{current}} \quad (4)$$

The targeted energy performance benchmarking values are median, average, upper hinge, upper whisker and lower hinge that derived from the energy performance boxplot of each building type. Specifically, lower whisker is not used as benchmarking value as the low energy use buildings in various types might bias the calculation results. The buildings that currently consume less than the benchmarking values of corresponding types are considered with the saving potential of 0. The outliers are not removed in the energy saving potential calculation.

The statistical analysis is performed by using Python and validated in STATISTICA software. Tableau is used for data visualization. The detailed results are presented in the next Section.

3. Results

3.1. Descriptive statistics

Table 1 presents the descriptive statistics results of the public building database, including the total number of buildings and floor area of each building type (in percentage), as well as the average and standard deviation of the key parameters: building age, useful floor area and measured energy use.

Table 1. Descriptive statistics of Flemish public building database

Category	**Type**	**Count [%]**	**Total Floor Area [%]**	**Building Age [Year]**		**Useful Floor Area [m2]**		**Measured Energy Use [kWh/m2/y]**	
				σ	**μ**	**σ**	**μ**	**σ**	**μ**
Educational	Daycare and/or after school care	3.4%	0.8%	57	52	893	889	230	117
	Pre-primary school	4.0%	1.3%	65	54	1211	992	188	95
	Primary school	21.8%	12.9%	75	44	2192	1715	172	139
	Secondary school	9.4%	21.0%	76	43	8338	6984	165	116
	Higher education and universities	1.8%	5.7%	81	103	11465	12670	229	124
	Other educational infrastructure	4.6%	4.2%	82	80	3367	3440	220	165
Office	Administrative building	13.9%	12.5%	80	105	3332	5725	242	171
Healthcare	Hospital	1.6%	10.7%	58	35	25074	22011	393	131
	Elderly home	6.2%	10.2%	41	33	6084	3788	333	193
	Other welfare provision	5.2%	5.2%	62	74	3678	7370	254	194
Sports	Sports hall with swimming pool	0.7%	1.1%	44	17	5382	3279	978	1953
	Swimming pool	1.0%	0.6%	42	20	2480	1795	1146	469
	Sports hall	6.0%	4.0%	38	21	2495	2413	270	249
Cultural events	Cultural or meeting building	11.1%	4.6%	82	98	1551	2755	223	153
	Museum	1.4%	1.2%	205	206	3123	4224	271	156
	Library	2.7%	1.4%	55	63	1977	4342	229	154
Public services	Station building	0.6%	0.3%	82	45	1882	3554	490	222
	Airport building	0.02%	0.1%	70	27	9912	4209	888	28
	Police office	1.3%	0.9%	62	64	2774	4552	299	123
	Post office	3.1%	1.0%	43	40	1198	3746	300	105
	Justice court	0.2%	0.2%	90	50	3495	4871	208	55

In total there are 9141 valid buildings with total floor area of around 3.4x10^7 m^2. 21 building types are further grouped into 6 major building categories based on their functionalities. More than 40% in the database are educational buildings, following by administrative offices and buildings for cultural events. Similarly, educational buildings have the largest total floor area of more than 40%, following by the floor area of healthcare buildings of around 25%.

3.2. Clustering results

The *k*-mean method is applied to 21 types of buildings in the EPC public building database respectively. Table 2 summarizes the optimal *k* values, *SC* values and the details of clusters with two data standardization methods. With z-score method, in total, 76 clusters are formed, and the number of clusters of each building type individually varies from 2 to 6. Two airport buildings are not clustered. The *SC* values are generally well above 0, ranging from 0.28 to 0.85, which gives a relatively dense clustering result. For instance, in educational category, 1962 primary school buildings, with outliers identified and excluded, are formed into 4 clusters, 147 in Cluster 1, 1122 in Cluster 2, 62 in Cluster 3 and 631 in Cluster 4. With min-max method, 52 clusters are totally formed, and the number of clusters of each building type varies from 2 to 4. The *SC* values range from 0.32 to 0.61. Larger SC values are obtained in 17 (out of 21) building types by using min-max method, which indicates a comparably better clustering result on average. Similarly, for primary school buildings, 2 denser clusters are formed with 707 in Cluster 1 and 1255 in Cluster 2. It can be concluded that different data standardization methods may result in different clusters for the same building type, and some building types could be potentially merged based on their similar cluster results. Thereafter, archetypes can be represented by their corresponding cluster centroids, with the representative construction year, useful floor area and measured specific energy use of each cluster.

Table 2. Clustering results of Flemish public building database

Category	**Type**	**Z-score method**			**Min-Max method**		
		k	***SC***	**Clusters**	***k***	***SC***	**Clusters**
Educational	Daycare and/or after school care	5	0.39	C1: 129, C2: 88, C3: 32, C4: 7, C5: 35	2	0.46	C1: 191, C2: 100
	Pre-primary school	4	0.36	C1: 26, C2: 68, C3: 209, C4: 58	2	0.48	C1: 276, C2: 85
	Primary school	4	0.37	C1: 147, C2: 1122, C3: 62, C4: 631	2	0.52	C1: 707, C2: 1255
	Secondary school	4	0.37	C1: 10, C2: 491, C3: 206, C4: 107	3	0.45	C1: 91, C2: 504, C3: 219
	Higher education and universities	4	0.44	C1: 38, C2: 15, C3: 9, C4: 99	3	0.48	C1: 108, C2: 15, C3: 38
	Other educational infrastructure	4	0.39	C1: 204, C2: 62, C3: 4, C4: 128	3	0.44	C1: 129, C2: 210, C3: 59
Office	Administrative building	5	0.48	C1: 784, C2: 74, C3: 8, C4: 314, C5: 22	2	0.61	C1: 324, C2: 878
Healthcare	Hospital	2	0.33	C1: 44, C2: 93	4	0.32	C1: 50, C2: 10, C3: 40, C4: 37
	Elderly home	4	0.34	C1: 176, C2: 290, C3: 4, C4: 58	2	0.46	C1: 196, C2: 332
	Other welfare provision	5	0.50	C1: 296, C2: 116, C3: 6, C4: 3, C5: 32	2	0.57	C1: 312, C2: 141
Sports	Sports hall with swimming pool	4	0.32	C1: 16, C2: 10, C3: 22, C4: 2	3	0.35	C1: 12, C2: 16, C3: 22
	Swimming pool	2	0.28	C1: 28, C2: 43	2	0.45	C1: 60, C2: 11
	Sports hall	6	0.43	C1: 185, C2: 285, C3: 17, C4: 34, C5: 1, C6: 2	2	0.53	C1: 322, C2: 202
Cultural events	Cultural or meeting building	5	0.47	C1: 616, C2: 43, C3: 224, C4: 2, C5: 48	2	0.59	C1: 689, C2: 244
	Museum	4	0.51	C1: 79, C2: 20, C3: 12, C4: 4	4	0.55	C1: 83, C2: 19, C3: 9, C4: 4
	Library	2	0.83	C1: 218, C2: 2	2	0.61	C1: 173, C2: 47
Public services	Station building	4	0.43	C1: 29, C2: 1, C3: 11, C4: 12	4	0.49	C1: 31, C2: 11, C3: 10, C4: 1
	Airport building	-	-	-	-	-	-
	Police office	2	0.43	C1: 87, C2: 20	2	0.53	C1: 87, C2: 20
	Post office	2	0.85	C1: 268, C2: 2	2	0.45	C1: 126, C2: 144
	Justice court	4	0.44	C1: 9, C2: 3, C3: 2, C4: 5	4	0.44	C1: 5, C2: 10, C3: 3, C4: 1

3.3. *Energy performance benchmarking and energy saving potential*

Table 3 summarizes the current energy use of each building category, as well as their relative energy saving potential (in percentage) and the corresponding total floor area under different benchmarking values. Buildings that perform worse than the benchmarking value are considered in the corresponding energy saving potential calculation, and that specific benchmarking value is used as the new energy performance while calculating the saving potential. Within one building category, the saving potential increases in the order of upper whisker, upper hinge, average, median and lower hinge. Statistically, when only focusing on the least energy efficient buildings identified as upper whisker, the saving potential largely comes from educational buildings - mainly due to their large quantity in the database, and sports activities buildings - mainly due to their high current energy consumption. Buildings for public services show less relative saving potential at all benchmarking levels because of its limited number and better energy performance. With altering the benchmarking value to a more ambitious level, the saving potential gradually increases in all building categories, and buildings for cultural events stand out with a higher relative saving potential. With the most ambitious "lower hinge" target, relative energy saving potential ranges from 27.4% of buildings for public service, up to 43.2% of buildings for cultural events, and the actual energy savings, more than 65%, are mainly from educational and healthcare buildings, due to the large weight of these building categories in the database. Current energy consumption could be reduced by 33.6%, which amounts to more than 2.9 TWh in total in Flanders.

Table 3. Relative energy saving potential and the corresponding floor area with different benchmarks

Category	**Current**	**Upper Whisker**		**Upper Hinge**		**Average**		**Median**		**Lower Hinge**	
	[MWh]	**[%]**	**10^5*[m^2]**	**[%]**	**10^5*[m^2]**	**[%]**	**10^5*[m^2]**	**[%]**	**10^5*[m^2]**	**[%]**	**10^5*[m^2]**
Cultural events	609.8	5.6	1.0	15.1	7.1	21.8	10.9	27.7	13.7	43.2	20.2
Educational	2797.1	4.9	7.6	12.9	36.9	16.8	53.4	21.8	74.6	33.8	116.4
Healthcare	3180.2	2.1	2.0	9.0	26.9	15.9	46.2	17.6	50.6	31.1	70.1
Office	1015.8	4.3	2.2	12.2	10.5	16.3	15.6	21.2	21.1	34.7	32.7
Public service	282.0	0.4	0.2	6.3	2.5	12.5	4.1	15	4.9	27.4	7.0
Sports	862.4	11.1	0.6	15.7	4.1	17.8	5.8	24.3	8.7	36.5	14.2
Total	8747.3	4.3	13.5	11.6	87.9	16.7	136	20.7	173.6	33.6	260.6

Furthermore, Figure 1 provides a more in-depth description on the relative energy saving potential under different benchmarking values for 6 building categories at the provincial level. The relative energy saving potential of specific building category could be identified separately for five different Flemish regions (Antwerp, East Flanders, Flemish Brabant, Limburg and West Flanders).

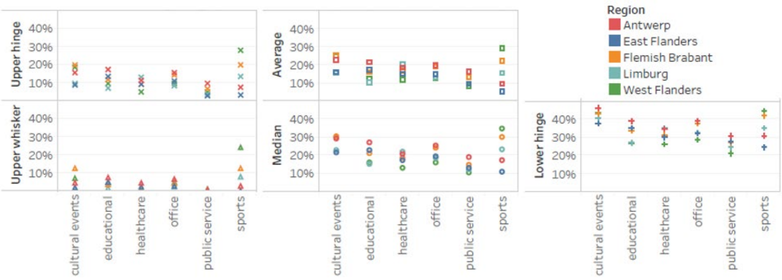

Figure 1. Relative energy saving potential with different benchmarks per category per region

Moreover, Figure 2 shows an example hotspot map of public building relative energy saving potential, where median is used as the benchmarking value.

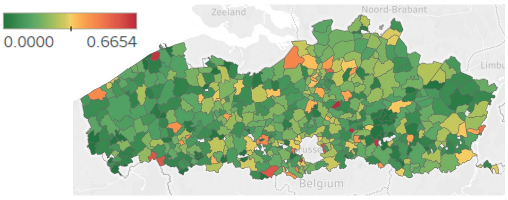

Figure 2. Hotspot map of public building relative energy saving potential

The relative energy saving potential, benchmarked by median, is up to 66.5% in certain regions and municipalities. Building sets in these regions should be given priorities in the renovation plan. Some missing places can be found in the map, indicating that public building EPC data of these places are unavailable in the current database yet. Moreover, it should be noted that similar hotspot maps can be generated by using relative potential on the basis of other benchmarking values, as well as by using the actual energy saving potential.

4. Conclusion and discussion

A cluster analysis is performed on around 9000 public buildings in Flanders. The clusters vary based on the data standardization methods. Some similar clusters could be merged potentially. The clustering results, together with statistics on thermal properties and technical building systems, will further serve as archetype inputs for the development of a bottom-up energy simulation model of the Flemish non-residential building stock. Extra parameters could be potentially included in future work by combining the EPC public database with other data sources (e.g. detailed geometry in 3D). It should be noted that the archetypes need to be diverse enough to represent the entire building stock and improve the accuracy of energy saving potential results. The energy saving potential with different benchmarking values is further calculated and grouped by building category and region. The result reveals clearly an untapped energy saving potential in the public building sector in Flanders, meanwhile, the energy inefficient building sets are identified and their saving potential are quantified stepwise. Furthermore, the energy saving potential could be investigated for the identified clusters in future work, and the poorly performing clusters could be further targeted in future renovation plan by relevant authorities.

Acknowledgement

The work is supported via the energy transition funds project 'EPOC 2030-2050' organised by the Belgian FPS economy, S.M.E.s, Self-employed and Energy.

References

[1] EUROSAI (2018). Energy Efficiency of Public Sector Buildings. Retrieved January 15, 2019, from https://www.eurosai.org/en/databases/audits/Energy-efficiency-of-public-buildings/

[2] Singh, M.K., Mahapatra, S. and Teller, J. (2013). An analysis on energy efficiency initiatives in the building stock of Liege, Belgium. Energy Policy, 62, 729-741. https://doi.org/10.1016/j.enpol.2013.07.138

[3] EPBD (2018). Directive (EU) 2018/844 of the European Parliament and of the Council of 30 May 2018 amending Directive 2010/31/EU on the energy performance of buildings and Directive 2012/27/EU on energy efficiency. Retrieved January 15, 2019, from http://data.europa.eu/eli/dir/2018/844/oj

[4] TABULA (2012). Typology Approach for Building Stock Energy Assessment. Retrieved January 15, 2019, from http://episcope.eu/iee-project/tabula/

[5] Mata, É., Sasic Kalagasidis, A. and Johnsson, F. (2014). Building-Stock Aggregation through Archetype Buildings: France, Germany, Spain and the UK. Building and Environment, 81, 270-282. http://dx.doi.org/10.1016/j.buildenv.2014.06.013

[6] Cerezo Davila, C., Sokol, J., Alkhaled, S., Reinhart, C., Al-Mumin, A., and Hajiah, A. (2017). Comparison of four building archetype characterization methods in urban building energy modeling (UBEM): A residential case study in Kuwait City. Energy and Buildings, 154, 321–334. http://doi.org/10.1016/j.enbuild.2017.08.029

[7] Ghiassi, N., and Mahdavi, A. (2017). Reductive bottomup urban energy computing supported by multivariate cluster analysis. Energy and Buildings, 144, 372–386. http://doi.org/10.1016/j.enbuild.2017.03.004

[8] EPC Publieke Gebouwen (2019). Retrieved January 15, 2019, from https://data.gov.be/

[9] Tukey, J.W. (1977). Exploratory Data Analysis. Addison-Wesley. ISBN 978-0-201-07616-5. OCLC 3058187. https://doi.org/10.1002/bimj.4710230408

[10] Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1), 100–108. https://doi.org/10.2307/2346830

[11] Peter J. Rousseeuw (1987). Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. Computational and Applied Mathematics, 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7