Statistical data-driven analysis and modelling of total energy use in new or thoroughly renovated single-family houses

Matthias Y.C. Van Hove[1*], Jelle Laverge[1], Arnold Janssens[1], Marc Delghust[1]
[1]Ghent University, Ghent, Belgium
[*]matthias.vanhove@ugent.be

Abstract

The subject of this paper is to analyse how the regulatory calculated energy use relates to the real total energy use for new or thoroughly renovated Flemish single-family houses where electricity is the only energy carrier. Additionally, the authors determine whether statistical data-driven models can help inform current and future home owners and tenants about their energy use (and thus also potential energy savings when applying energy saving measures). These questions are investigated by using housing datasets from the Flemish energy performance database and real energy use data from the Belgian grid operator. The paper comprises outlined database cleansing and filtering choices and enlightening statistical database analyses and figures. The results clearly demonstrate that the regulatory calculation method poorly estimates the real energy use (RMSE-/MAE-results of respectively 7227 kWh/y and 5242 kWh/y), yet both are moderately correlated ($\tau = .548, p < .001$). Further, the statistical regression models show good results at stock level for new or thoroughly renovated Flemish single-family houses (where electricity is the only energy carrier) (adj. R^2 up to 65.3%). Nevertheless, their performance at individual building level is still limited and considered too poor for inference as a considerable part of the variance is left unexplained.

Keywords

Energy Performance database
Data-driven regression modelling
real yearly building energy use
Statistical data analysis

Introduction

The building sector is responsible for approximately 40% of EU's energy consumption and 36% of the CO_2-emissions and is therefore considered the largest energy consumer in Europe (EU, 2020). In order to reduce the primary energy consumption and CO_2-emissions of buildings at stock level by 2050, the EU in 2002 introduced official energy performance regulations for new and existing buildings undergoing major renovation (EU, 2003). In Flanders, these European guidelines from the Energy Performance of Buildings Directive (EPBD) were implemented in the Flemish Energy Performance and Indoor Climate Decree (EPB decree) (Flemish Authorities, 2009). Since 2006, it requires every newly built or thoroughly renovated house to meet the official energy performance requirements (*i.e.*, *i.a.* predefined building energy performance level (E-level)).

The energy performance level in Flanders is calculated using standardised, simplified calculation procedures (VEKA, 2019), based on building characteristics and a standard average user profile (CA EPBD, 2013) (as is the case in other countries). The question however often arises to what extent the regulatory calculated energy uses and performance indicators relate to real energy uses and if calculated energy savings associated with better building performance levels are fully obtained in practice?

Statistical studies in other countries (*i.e.*, The Netherlands (Majcen *et al.*, 2013), Germany (Sunikka-Blank *et al.*, 2012), France (Cayre *et al.*, 2011), UK (Kelly e*t al.*, 2011) and Switzerland (Cozza *et al.*, 2020), which investigated the accuracy and outcomes of their national implementation of the European calculation method, indicated that, on average and especially for less energy-efficient dwellings, the regulatory energy calculations tend to overestimate the total building energy use. For new or thoroughly renovated energy efficient buildings, an underestimation of the total building energy use is recognised (Cozza *et al.*, 2020; Sunikka-Blank *et al.*, 2012).

In a study conducted in Belgium on a small sample of new, low-energy houses (Delghust *et al*., 2015), researchers found that the total, real energy use was slightly underestimated by the regulatory EPB calculations (most likely due to the electricity use for lighting and domestic appliances which are not considered in the EPB regulation). On the contrary, the natural gas consumption for space heating (SH) and domestic hot water (DHW) was strongly overestimated by the calculation method (*i.e.*, by on average 25%).

The work reported in this paper aims to answer the above questions on the relationship between the regulatory calculated and real energy use, specifically for the total energy use in recently built or thoroughly renovated single-family houses in Flanders that only

have electricity as an energy carrier. Further it aims to determine whether data-driven statistical methods (*i.e.*, classical linear regression) based on building characteristic data and supplemented with end results from the energy performance calculation have better performance, both at stock level as well as at individual building level. The work reported is part of a broader study in collaboration with the Flemish Energy and Climate Agency in which both the natural gas and electricity consumption of modern and old Flemish single-family houses was studied and compared with the predictions from the regulatory energy calculation (Van Hove *et al*., 2021).

Materials and methods

Data Overview

The study focuses on modern single-family houses (*i.e.*, built or thoroughly renovated since 2006). All cases are acquired from the Flemish Open Data Portal (OpenData Vlaanderen, 2020). The building characteristic data is gathered directly from the Open Data Portal, supplemented with missing technical system data and detailed building geometry data from the underlying energy performance database of the Flemish Energy and Climate Agency (VEKA) and real energy use data and installed PV power figures from the Belgian grid operator.

The analysis focused on *(i)* the relation between real energy use and common building characteristics and *(ii)* the performance of statistical data-driven models to predict the real energy use based on common building characteristics.

The scope of the study was limited to single-family houses (*i.e.*, no apartments). In order to have reliable figures of the total actual energy consumption, only homes were selected with (1) an individual heating system, (2) where space heating and DHW are only on gas and/or electricity and (3) that were occupied and for which meter data was available for more than one year (*i.e.*, no default or estimate values from the energy utilities based on previous meter data). Both for gas and electricity, normalised annual consumption figures for each single-family house were obtained for the years 2012-2019.

The EPB-registry entries had to be free from any major error or shortcoming with regard to the data (*e.g.*, missing data or contradictions). The EPB registry is one centralised database with data from the official EPB-calculation files of all new and thoroughly retrofitted buildings (*i.e.*, residential buildings but also offices, schools *etc.*). The Open Data Portal then is an open online platform which makes part of the EPB-data, anonymised, publicly available. It however only contains some of the most important variables (*e.g.*, the building size, average insulation levels, present technical systems) as well as intermediate and final results of the energy performance calculation. From this database a preselection of 135166 cases was gathered based on the study's focus on new or thoroughly renovated single-family houses.

Data Treatment

Data-cleansing (based on analysis within dataset and across the available datasets) revealed a substantial number of contradictions, indicating errors between data-files, changes to the dwelling after the moment of completion and EPB assessment or missing data that requires precautions with regard to other variables or even the cases themselves. For example, whereas the EPB database indicated the presence of PV panels in 21% of the houses, more recent energy utility data indicated the number increased to 33%. Further, 29% of the available cases only had natural gas consumption figures but no electricity consumption figures, which meant that these cases could only be used for the analyses on natural gas consumption but not for analyses on total energy consumption. Disputable data and all derived variables that could not be corrected were marked as missing data and excluded from further analysis and statistical modelling.

The reliability of the real consumption data was also an important filtering criterion. For gas and electricity, precise consumption figures were available through the annual meter readings. However, for bulk energy resources (wood, pellets, coals, fuel oil and gas cylinders) no such accurate data were available. Therefore, 27% of the cases were excluded from the final subset on the space heating and DHW energy use because they used bulk energy resources in addition to natural gas. Furthermore, cases with electricity consumption figures of 0 kWh/y were also excluded from the final subset since negative real energy consumption figures are not being reported by the energy utilities (*i.e.*, they are set to zero), which could lead to a mismatch in statistical models.

Additionally, a small percentage of cases had a significantly larger real energy use compared to the other cases in the dataset (*i.e.*, very skewed distribution at the high end). Because the top 2.5% of natural gas consumption figures were on average 16 times higher than the median natural gas consumption and the top 2.5% of electricity consumption figures were on average 11 times higher than the median electricity consumption, it was decided to consider the top 2.5% natural gas and electricity consumption figures as outliers and exclude them from further analysis and statistical modelling. For comparison, the 95-percentiles for both natural gas and electricity consumption were only 2 and 3 times higher than the median energy consumption figures. After data treatment, cleansing and filtering, the final dataset (O) comprised 68228 cases which corresponds to 50.5% of the cases in the originally received dataset.

Descriptive Statistics

The total Flemish single-family housing stock included 2,150,000 single-family houses in 2018 (VEKA, 2018). The final subset (after cleaning, filtering and coupling) that is studied in this paper therefore represents ~3.2% of the total single-family housing stock and 50% of the total EPB-rated single-family housing stock (VEKA, 2018). As can be seen in *Figure 1*, most of the single-family houses in the EPB registry are rated E61-80.

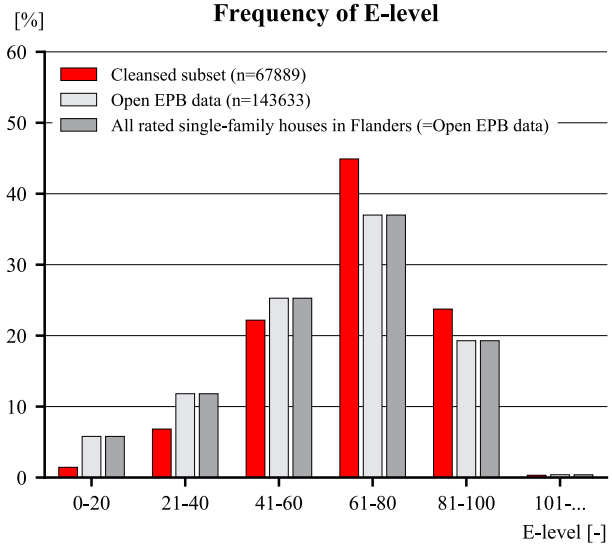

Figure 1: Shares of energy rated cases in the studied sample, the Open EPB dataset and the Flemish single-family housing stock (2018).

According to the VEKA and Statbel (Statbel, 2020/2020), 42% of Flemish single-family houses are detached houses, 27% are semi-detached houses and 30% are terraced houses (*Figure 2*). Our final sample of single-family houses has less terraced houses and more semi-detached houses than the total Flemish single-family housing stock.

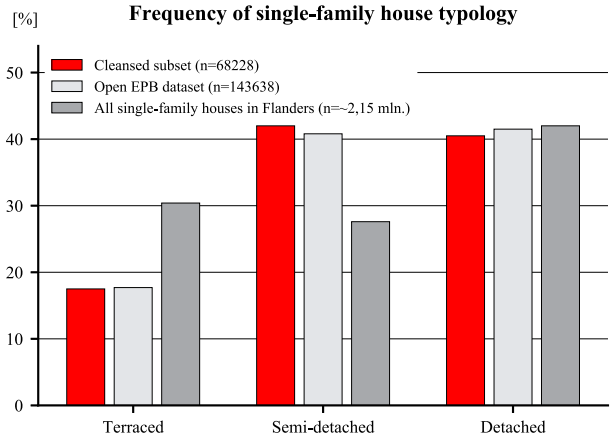

Figure 2: Shares of single-family houses per typology.

Categorisation

The final dataset comprised single-family houses with all possible combinations of technical and renewable energy systems (*e.g.*, solar panels, condensing boiler, ventilation system) and energy carriers (*i.e.*, gas and electricity) for different end uses (*e.g.*, space heating, domestic appliances, indoor air quality, DHW). The amount of collected data made it possible to define seven categories, each containing enough cases for statistically valid analysis (*Table 1*).

Table 1: with categories O1 and O2 applying to all houses in the cleansed dataset, categories A2, A3, A4 and A5 applying to all houses with natural gas for space heating and/or DHW (and cooking) and category B1 applying to all houses with electricity as the only energy carrier.

Category	ANALYSIS	
	Calculated energy use	Measured energy use
O1	Total	nat. gas and electricity
O2	aux. + solar + cooling	electricity
A2	space heating + DHW	nat. gas
A3	space heating	nat. gas
A4	DHW	nat. gas
A5	/	nat. gas (only cooking)
B1	Total	electricity

This paper focuses only on the analyses and statistical models for category B1, namely analyses concerning the fit and the prediction error regarding the total yearly energy use for single-family houses that only have electricity as an energy carrier (*i.e.*, the houses do not have other energy carriers). Category B1 comprises 1947 cases and evidently contain single-family houses where the majority has a heat pump and/or PV-panels. Note that the real electricity consumption figures also include electricity consumption for domestic appliances and cooking, which is not included in the regulatory calculated energy uses.

Data pre-processing for statistical modelling

After data-cleansing, extra pre-processing steps were needed to make the data ready for statistical modelling. To be able to include categorical variables in statistical models, one-hot encoding was used to translate them to dummy variables.

Moreover, the data underwent feature scaling to normalise the range of the independent variables. For this feature scaling a robust scaling (*1*) was preferred since it performs better in case of outliers in the data (*N.B.*, the features contained outliers).

$$X_{rs} = \frac{X_i - Q_2(X)}{IQR(X)} \quad (1)$$

with $Q_2(X)$ the median of the explanatory variable and $IQR(X)$ the interquartile range of variable X (*i.e.*, $IQR(x) = Q_3(x) - Q_1(x)$).

Then, the dataset was divided into a training set and a test set by applying an 80%-20%-ratio. The training set was used once to fit the model. The test set was then used to make an objective evaluation of the fit of the model with unseen input data and determine the level of model generalisation.

Statistical Methods and Models

Statistical analyses on the data are all conducted in Python with the Pycharm IDE and the statistical packages 'scikit-learn' (Pedregosa *et al.*, 2011) and 'statsmodels' (Skipper *et al.*, 2010) in combination with the data analysis and visualisation package 'pandas' (McKinney, 2010). All reported correlation values were obtained using the non-parametric Kendall's Tau rank correlation since it does not rely on any underlying distribution for the analysed variables. A p -value of .05 was used for null hypothesis significance testing (two-tailed) (*i.e.*, did the result occur due to chance or not).

The data-driven statistical modelling technique that was tested and compared with the regulatory energy calculation (EPB) is a classical linear regression. In this linear regression, the ordinary least squares (OLS) estimation method was used to generate unbiased regression coefficients.

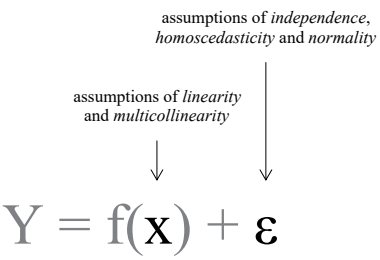

Figure 3: Overview of necessary assumptions for linear regression models.

In order to fulfil the necessary assumptions for linear regression models, the model input variables were checked for linearity, autocorrelation and multicollinearity; the residuals were checked for independency, homoscedasticity and normality (*Figure 3*).

As discussed above, all model input variables were normalised. As a result, the magnitude of the regression coefficients gives an indication of the parameters relative importance in the regression model. In order to evaluate the obtained regression models and compare them to the official energy calculation method, additional model performance metrics were used such as adjusted R-Squared, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).

Results

Comparison of real energy use per E-level

In order to understand how the E-level relates to the prediction error between the real and regulatory calculated energy consumption, the total energy use is examined for different E-level categories. In *Figure 4*, the annual real and regulatory calculated primary total energy uses, for dwellings where the only energy carrier is electricity (*i.e.*, cluster B1 in *Table 1*), are shown and in *Figure 5*, the primary total energy use per square meter of floor area [kWh/m² •y] is shown.

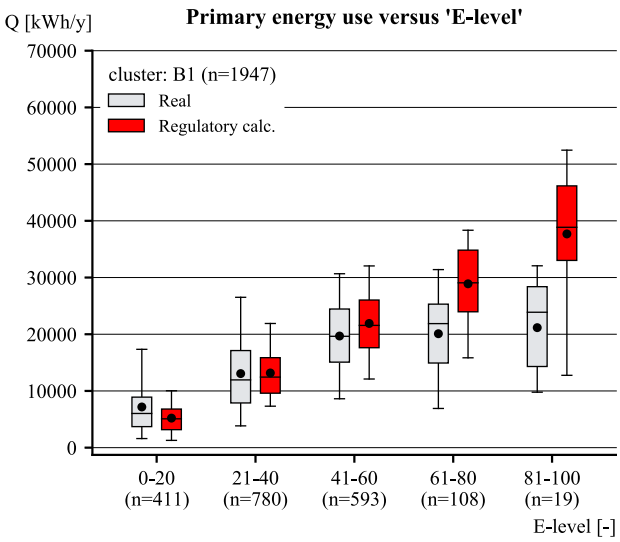

Figure 4: Boxplot of the regulatory calculated annual primary total energy use and the real annual electricity consumption in dwellings with electricity as only energy carrier in function of the E-level. The boxplots show 5%, 25%, 50%, 75% and 95%-percentiles as well as a mean value (i.e., black dots).

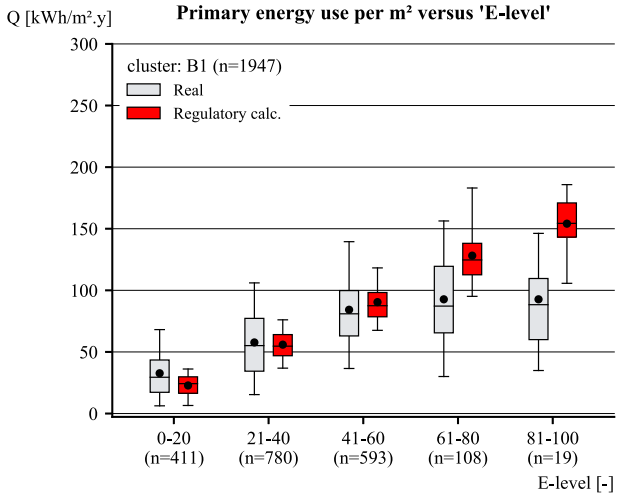

Figure 5: Boxplot of the regulatory calculated annual primary total energy use per m2 of floor area and the real annual electricity consumption per m2 in dwellings with electricity as only energy carrier in function of the E-level. The boxplots show 5%, 25%, 50%, 75% and 95%-percentiles as well as a mean value (i.e., black dots).

From these figures it is clear that there is a strongly positive correlation ($\tau = .548, p < .001$) between the E-level and the real electricity consumption (*i.e.*, the higher the E-level, the higher the real electricity consumption of the dwelling). Nevertheless, there is a clear difference between the regulatory calculated and real electricity consumption for each E-level category. For the better E-levels (*i.e.*, E0-E40), the regulatory calculation method clearly underestimates the real energy use of the dwellings, whereas for the poorer E-levels (*i.e.*, E41-E100), the regulatory calculation method overestimates the real energy use. The real electricity consumption [kWh/y] is overestimated in the

EPB calculation method by on average 21% (*i.e.*, the average prediction error, which is the regulatory calculated energy consumption minus the real energy use as a percentage of the real energy use). The spread of relative prediction error across E-levels (*i.e.*, from 110% underestimation to 71% overestimation) indicates a lack of fit.

EPB calculation method performance

The poor fit between the annual real and regulatory calculated electricity consumption is further demonstrated in *Figure 6*. In an ideal scenario, a linear function $y = x$ should closely describe the relationship between both variables. However as expected and based on earlier findings (above), this ideal relationship is not obtained, although a clear trend is visible.

The R-Squared model performance for the 1 on 1 comparison (*Figure 6*) shows that 20.2% of the variance in real electricity consumption was predicted by the EPB-calculation method. RMSE- and MAE-results of respectively 7227 kWh/y and 5242 kWh/y however show that the accuracy of the predictions is rather low. Note that the regulatory calculated primary energy consumption in *Figure 6* is the output of the EPB calculation method (so not yet from a linear regression model).

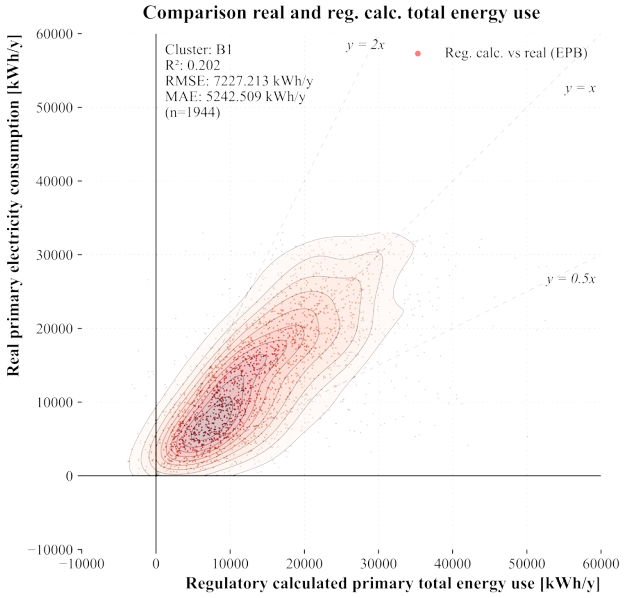

Figure 6: Scatter plot of the regulatory calculated annual primary total energy use and the real annual electricity consumption in buildings with only electricity as energy carrier.

Statistical Data-driven Modelling

In this section, the performance of the official EPB calculation method is compared to data-driven linear regression models in three steps. (1) First, a simple regression model was tested with the regulatory calculated energy consumption as single explanatory variable (*N.B.*, maybe a simple scaling of the regulatory calculated energy use already has good fit with the real energy uses). (2) Secondly, a more black-box multiple linear regression model was tested with common building characteristics (*i.e.*, common building features for potential home owners and tenants, thus excluding building energy performance calculation figures and detailed building characteristics (*e.g.*, system efficiency, thermal resistance of the envelope *etc*.)). (3) Thirdly, a multiple linear regression model was tested with all available parameters from the EPB-database (excluding those suffering from problems with multicollinearity or autocorrelation). This three-step approach allows to demonstrate whether and which statistical models of increased complexity can help predict real yearly building energy use and possibly replace the current regulatory monthly method. Note that for all models the residuals were inspected (*i.e.*, model diagnostics) to ensure the quality of the model.

The validation output of the first simple linear regression model with the regulatory calculated energy use as only explanatory variable resulted in a RMSE and MAE of respectively 5009 kWh/y and 4006 kWh/y. This indicates that a simple scaling and shifting of the regulatory calculated energy figures already gives significantly more accurate results as compared to the pure regulatory calculated energy uses. The R-Squared model performance of 53.1% shows that more of the variance in real electricity use is explained by the model, yet the performance is still not great. The regression coefficients and (bootstrapped) confidence intervals of model 1 are given in *Table 2*.

Table 2: Simple regression model output (n=1944). Note that the regression coefficients and 95% CI are normalised coefficients.

REGRESSION MODEL 1			
	β	95% CI	*p*
(constant)	13311.4	[12832.4, 13692.2]	<.001
EPB calc.	7955.2	[7517.4, 8668.3]	<.001

In the second multiple linear regression model, the real building energy use is predicted based on common building characteristics (*i.e.*, building features that inhabitants can easily fill in themselves to evaluate whether such a prediction model can be used in an open online regulatory tool for yearly energy use prediction). The validation output of the multiple linear regression model with common building characteristics as explanatory variables showed that 60.7% of the variance in real total energy use is predicted by common building parameters. The RMSE and MAE are respectively 4938 kWh/y and 3926 kWh/y, which is slightly better than the results for the simple model 1. The regression coefficients and confidence intervals of model 2 are given in *Table 3*.

Table 3: Multiple regression model output (n=1944) based on common building parameters. Parameter ending with '?' are Boolean parameters indicating if true or false as 1 or 0. Note that the regression coefficients and 95% CI are normalised coefficients.

REGRESSION MODEL 2

	β	95% CI	*p*
(constant)	17972.5	[17591.7, 18625.9]	<.001
Floor area	1252.4	[409.0, 1973.9]	<.001
Building volume	1666.5	[930.2, 2540.5]	<.001
Detached?	1178.7	[655.9, 1779.3]	<.001
Basement?	1017.8	[376.8, 1479.7]	<.001
EPB-cert. year	-2438.5	[-3368.0, -2003.1]	<.001
PV?	-8667.5	[-9511.1, -8247.2]	<.001
Number PV-panels	-448.0	[-731.8, -45.7]	<.001
Vent. system D?	-1294.7	[-1746.2, -546.2]	<.001
Space cooling?	1637.9	[576.9, 2355.6]	<.001

In the third multiple regression model, the modelling approach of the second model was repeated, but this time with all available EPB-parameters in the hope that more variance in the real energy use is explained by a statistical model based on common and detailed building parameters. The list of all available parameters is then reduced based on Kendall-Tau correlation results (with p <.05) of explanatory variables with the dependent variable (*i.e.*, the real annual natural gas consumption), the Variable Inflation Factor (VIF) results and by looking at the p-values and bootstrapped confidence intervals (CI) in the multiple linear regression output. The exclusion of features from the model is done one by one stepwise.

The validation output of the third multiple regression model demonstrated that a maximum of 65.3% of the variance in real energy use is explained by linear correlations with the available parameters in the EPB-database (avoiding possible autocorrelation and multicollinearity problems). The RMSE and MAE are respectively 4462 kWh/y and 3527 kWh/y which is slightly better than the results from both previous regression models. The regression coefficients and confidence intervals of model 3 are given in *Table 4*.

Table 4: Multiple regression model output (n=1944) based on common building parameters. Note that the regression coefficients and 95% CI are normalised coefficients.

REGRESSION MODEL 3

	β	95% CI	*p*
(constant)	18003.3	[17583.3, 18545.8]	<.001
EPB-calc. SH	1660.7	[1342.4, 2251.4]	<.001
EPB-calc. DHW	849.0	[537.6, 1335.0]	<.001
EPB-calc. PV	-2662.9	[-4016.9, -2233.9]	<.001
Floor area	1101.1	[449.1, 1430.5]	<.001
Basement?	635.1	[90.7, 1118.9]	<.001
EPB-cert. year	-1783.9	[-2415.0, -1311.9]	<.001
PV?	-5928.3	[-6587.6, -4663.3]	<.001
Number PV-panels	-270.3	[-452.0, -3.6]	<.001
Space cooling?	1428.0	[251.7, 2053.7]	<.001
Av. U-value glass	1474.0	[850.0, 2001.0]	<.001
Window surface	1222.7	[883.3, 1718.0]	<.001

Regression diagnostics for linear regression models assured the validity of the models, checking for possible autocorrelation and multicollinearity problems. For brevity, not all residual plots are presented. As an example, a Q-Q plot of the residuals of regression model 2 is shown in *Figure 8* as well as a plot of the residuals versus fitted values in *Figure 7*. Inspection of the Q-Q plots of the residuals show that the residuals are nearly normally distributed except for some outliers at both ends and that they are linear over a wide range of values. Furthermore, the residuals versus fitted values indicate that the residuals are nearly uncorrelated to the fitted values. Therefore, the assumptions for homoscedasticity and normality hold true.

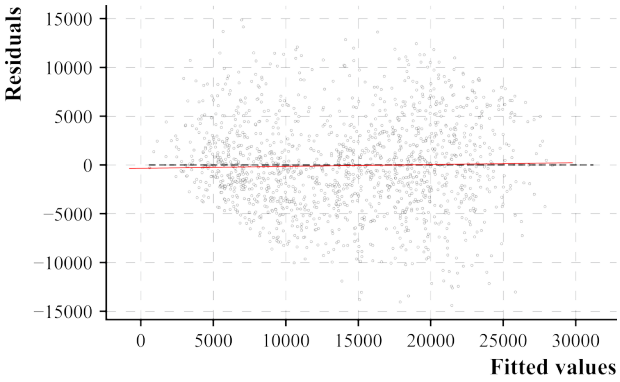

Figure 7: Plot of fitted values against residuals for regression model 2.

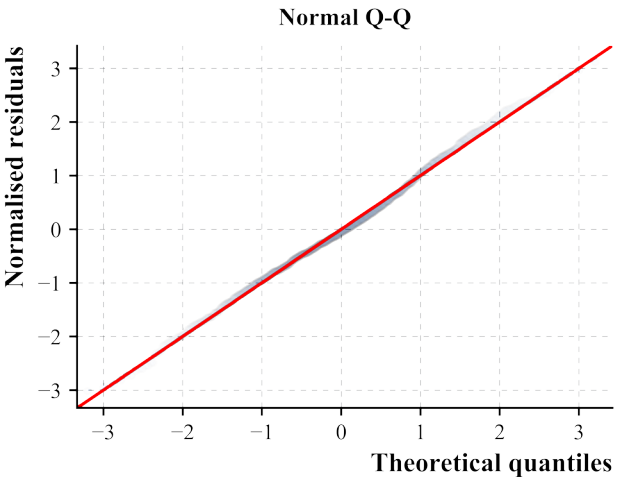

Figure 8: Normal Q-Q plot of the normalised residuals of the validation set for regression model 2.

An overview of the 1 on 1 comparison and the three studied regression models is given in *Table 5*. The regression models explain between 53% and 65% of the total building energy use. This is considerably higher compared to other studies in Germany (Rehdanz, 2007), Great Britain (Meier *et al.*, 2010), The Netherlands (Brounen *et al.*, 2012) and Switzerland (Cozza *et al.*, 2020), which explained 18 to 40% of the total energy use.

Table 5: Overview of model performance results from the EPB calculation method and data-driven linear regression models.

	adj. R2	RMSE	MAE
pure EPB-calc.	~20%	7227 kWh/y	5242 kWh/y
model 1	~53%	5009 kWh/y	4006 kWh/y
model 2	~60%	4938 kWh/y	3926 kWh/y
model 3	~65%	4462 kWh/y	3527 kWh/y

Conclusion

This study investigated the relationship between the real annual total energy use and the regulatory calculated energy use for single-family houses in Flanders with electricity as the only energy carrier. The real total energy use (or electricity consumption) is overestimated by the regulatory EPB-calculation method for the poorer E-levels and underestimated by the regulatory EPB-calculation for the better E-levels (+21% on average and RMSE-/MAE-results of respectively 7227 kWh/y and 5242 kWh/y). The lack of fit between both variables is confirmed by a largely variable prediction error among the E-level categories (*i.e.*, from 110% underestimation to 71% overestimation). Nevertheless, strong correlations are found between the annual real total energy use and the regulatory calculated primary total energy use ($\tau = .548, p < .001$). Additionally, a small but positive R-Squared value (*i.e.*, coefficient of determination) shows that part of the variance in the real electricity consumption is explained by the EPB calculation method which confirms that they are related (adj. R^2 = ~20%).

The output of a simple regression model with the regulatory calculated energy use as single explanatory variable immediately proved to have a significantly improved fit with the real total energy use as compared to the outputs of the regulatory calculation method (adj. R^2 = ~53%). Thus, through a simple scaling and shifting of the output of the regulatory calculation method, a much better estimation of the real total energy use is achieved. Data-driven (black-box) statistical regression models with common building characteristics and more detailed building characteristics show even better results with respectively 60.7% and 65.3% of the variance explained.

The results show that the performance of statistical regression models is decent compared to similar research studies in other countries. Also, the models show promising results for predictions at stock level. Yet, for inference at individual building level, the performance is still too poor. Still short of 70-80% variance explained, a considerable part of the variance in annual total energy use is left unexplained. This means that a part of the variance in the total energy use has to be attributed either to parameters that are not listed among the variables of the EPB registry (*e.g.*, number of inhabitants, occupant behavior, appliance ownership, income, regional weather differences) or that the values of the parameters listed are inaccurate (*e.g.*, inaccurate standard values).

Acknowledgements

The authors want to thank the Flemish Energy and Climate Agency (VEKA) and the Belgian grid operator Fluvius for data collection and helpful feedback during the course of the study.

References

EU. Energy performance of buildings directive, European Commission Department of Energy, 2020.

EU, Directive 2002/91/EC of the European parliament and of the council (Directive No 2003/91/EC) (p. L 1/65-71), European Parliament and Council, 2003.

Flemish Authorities, *Decreet houdende algemene bepalingen betreffende het energiebeleid,* Belgisch Staatsblad - Moniteur Belge, pp. 46145-46191, Brussels: Flemish Authorities, 2009.

VEKA, *Energiebesluit Bijlage V, 2019.*

Concerted Action EPBD. *Implementing the energy performance of buildings directive (EPBD). Featuring country reports 2012.* Porto, 2013.

D. Majcen, L. C. M. Itard, H. Visscher, "Theoretical vs. actual energy consumption of labelled dwellings in the Netherlands: Discrepancies and policy implications", *Energy Policy*, vol. 54, pp. 125-136, 2013.

M. Sunikka-Blank, R. Galvin, "Introducing the prebound effect: The gap between performance and actual energy consumption". *Building Research & Information*, vol. 40(3), pp. 260-273, 2012.

E. Cayre, B. Allibe, M.-H. Laurent, D. Osso, "There are people in the house! How the results of purely technical analysis of residential energy consumption are misleading for energy policies", in: ECEEE 2011 Summer Study – Energy Efficiency First Found, A Low-Carbon Society, 2011.

S. Kelly, D. Crawford-Brown, M. G. Pollitt, "Building performance evaluation and certification in the UK: Id SAP fit for purpose?", *Renewable and Sustainable Energy Reviews*, vol. 16, pp. 6861-6878, 2012.

S. Cozza, J. Chambers, C. Deb, J.-L. Scartezzini, A. Schlüter, M. K. Patela, "Do energy performance certificates allow reliable predictions of actual energy consumption and savings? Learning

from the Swiss national database", *Energy and Buildings*, vol. 224, 2020.
M. Delghust, W. Roelens, T. Tanghe, Y. De Weerdt, A. Janssens, "Regulatory energy calculations versus real energy use in high-performance houses". *Building Research & Information*, vol. 43(6), pp. 675-690, 2015.
Van Hove, M., Delghust, M., Janssens, A. (2021). Analyse naar de haalbaarheid van statistische modellen die energiegebruik in woningen kunnen voorspellen op basis van gebouwparameters. https://www.energiesparen.be/marktonderzoek
Open Data Vlaanderen, https://opendata.vlaanderen.be/organization/vlaams_energie_en_klimaatagentschap_veka
VEKA, Energy statistics - existing buildings in Flanders. Flemish Energy and Climate Agency (VEKA), 2018. https://www.energiesparen.be/energiestatistieken-bestaande-gebouwen-in-vlaanderen
Statbel, Kadastrale statistiek van het gebouwenpark, 2020, https://bestat.statbel.fgov.be/bestat/crosstable.xhtml?view=dade64a8-39c8-498e-9736-16a27ef618a0
Statistics Flanders, Woningvoorraad, Statbel, 2020, https://www.statistiekvlaanderen.be/nl/woningvoorraad
Pedregosa, "Scikit-learn: Machine Learning in Python", *JMLR*, vol. 12, pp. 2825-2830, 2011.
Skipper, Seabold, J. Perktold, "Statsmodel: Econometric and statistical modeling with python". *Proceedings of the 9th Python in Science Conference*, 2010.
McKinney, Data structures for statistical computing in python, *Proceedings of the 9th Python in Science Conference*, vol. 445, 2010.
K. Rehdanz, "Determinants of residential space heating expenditures in Germany". *Energy Economics*, vol. 29, pp. 167-182, 2007.
H. Meier, K. Rehdanz, "Determinants of residential space heating expenditures in Great Britain", *Energy Economics*, vol. 32, pp. 949-959, 2010.
D. Brounen, N. Kok, J. M. Quigley, "Residential energy use and conservation: Economics and demographics", *European Economic Review*, vol. 56, pp. 931-945, 2012.