Available online at www.sciencedirect.com

ScienceDirect

Energy Procedia 132 (2017) 897–902

Energy Procedia

www.elsevier.com/locate/procedia

11th Nordic Symposium on Building Physics, NSB2017, 11-14 June 2017, Trondheim, Norway

Explaining variability in metered energy use for similar buildings using Bayesian inference

Martin Heine Kristensen*, Steffen Petersen

Department of Engineering, Aarhus University, Inge Lehmanns Gade 10, 8000 Aarhus C, Denmark

Abstract

Typologically identical buildings may exhibit large differences in energy use due to various stochastic phenomena. In this paper, we present a study on how these phenomena can be explained by analyzing 1,050 observations of metered district heating energy use from a sample of 350 similar detached single-family dwellings located in 37 city districts (urban areas) in the city of Aarhus, Denmark. The results indicate that annual variations within the same buildings due to e.g. weather conditions account for only a minor proportion of the overall data variance (approx. 10%-20%). A larger proportion (approx. 25%-51%) is capsulated and explained by phenomena between the typologically identical buildings, probably due to the stochastic nature of occupant behavior. The largest proportion (approx. 30%-65%) of the data variance is explained by the district location, which suggests the presence of a socio-economic effect influencing the level of energy use between city districts.

Peer-review under responsibility of the organizing committee of the 11th Nordic Symposium on Building Physics.

Keywords: Metered building energy use; Data variability; Hierarchical modeling; Bayesian inference; Socio-economic effect

1. Introduction

It is rather obvious that actual energy use for building operation varies with the type of building (office, homes, retail, etc.), but even typologically identical buildings that share the same energy ratings exhibit large differences in energy use, as was shown in a recent data visualization video of approx. 28,000 Danish single-family dwellings [1].

* Corresponding author. Tel.: +45 23 73 77 18.
E-mail address: mhk@eng.au.dk

1876-6102
Peer-review under responsibility of the organizing committee of the 11th Nordic Symposium on Building Physics
10.1016/j.egypro.2017.09.709

The influence of occupant behavior is often ascribed considerable effect [2], but in reality, we know only very little about the mechanisms and phenomena that drive differences in energy use.

In this paper, we present a study showing how occupant behavior and other stochastic phenomena seem to manifest themselves on different aggregated levels, i.e. variations in energy use within the same buildings and between buildings of similar typology, as well as variations driven by the building location within the same city. The study is based on an analysis of metered district heating energy use for a large sample of similar detached single-family dwellings, all constructed within the same city in a period characterized by uniform building regulations. The data variance is modeled and decomposed in a hierarchical structure whereby the prevalence of the above-mentioned phenomena is analyzed. To do so, we employ a probabilistic approach using Bayesian mixed-effects modeling and regression including information about building energy use, age, location and floor area.

Nomenclature

y	Metered building energy use for heating [kWh/year].
x	Heated building floor area [m^2].
μ	Mean building energy use [kWh/year].
σ	Standard deviation of building energy use [kWh/year].
φ	Mean district energy use [kWh/year].
τ	Standard deviation of district energy use [kWh/year].
α	Intercept (const. effect) of district model [kWh/year].
β	Slope of district model [kWh/m^2/year].
θ	Mean of intercept value [kWh/year].
ω	Standard deviation of intercept value [kWh/year].
i	Indexing the observations from 1 to N_j.
j	Indexing the buildings from 1 to M_k.
k	Indexing the districts from 1 to L.
N_j	Number of metered data points from building j.
M_k	Number of buildings in district k.
L	Number of districts in dataset.

2. Method

2.1. Data

A random sample of 350 detached single-family dwellings, all constructed between 2008 and 2010 in the municipality of Aarhus, Denmark, and supplied by the public district heating network, was selected for analysis. All buildings in this construction period are expected to fulfill identical energy requirements given by the Danish Building Regulation (BR08) in force at the time of construction. The buildings were thus assumed to exhibit similar building physical properties with between-building variations being driven mainly by other factors than the building physical properties.

For each building, information about the construction year, heated floor area and location of the building within the city (urban area code) was collected from the publically available Building and Dwelling Register (BDR) that contains information about the Danish building stock. In total, 37 city districts (urban areas) were represented, each featuring between one and 60 of the 350 sampled buildings. Furthermore, the annual district heating energy use for the last three years was collected for all buildings yielding three observations per building – in total comprising 1050 observations. The district heating energy use consists of energy use for hydronic space heating in e.g. radiators and underfloor heating, and energy use for on-site domestic hot water (DHW) preparation.

2.2. Hierarchical regression model

A three-level mixed-effects model was assumed for the metered annual district heating energy use data, y_{ijk}, by fitting a hierarchical structure to observations $i = 1,..,N_j$ for each building $j = 1,..,M_k$ within the $k = 1,..,L$ city district groups:

Level one (Observations):

$$log(y_{ijk}) \sim N(\mu_{jk}, \sigma^2). \quad (1)$$

Level two (Buildings):

$$\mu_{jk} \sim N(\varphi_k, \tau^2). \quad (2)$$

$$\varphi_k = \alpha_k + \beta \cdot x_{jk}, \quad (3)$$

Level three (Districts):

$$\alpha_k \sim N(\theta, \omega^2). \quad (4)$$

At level one, the logarithm of the observed annual district heating energy use, $log(y_{ijk})$, was assumed to be independent and identically distributed (i.i.d.) random samples from a Gaussian distribution with unknown mean building energy use μ_{jk} and standard deviation σ. For simplicity, homoscedasticity was assumed across buildings such that $Var\left(log(y_{ijk})\right) = \sigma^2 \forall j$. At level two, the building mean energy use, μ_{jk}, was likewise assumed i.i.d. randomly sampled from a Gaussian distribution with unknown district means φ_k and homoscedastic district variations, i.e. $Var(\mu_{jk}) = \tau^2 \forall k$. By modeling the district means, φ_k, as a linear function (in the logarithmic domain) of the building floor area, x_{jk}, with a constant slope, β, and district-specific intercept, α_k, the effect of the individual districts can be inferred as the posterior variations in α_k. At level three, the intercept, α_k, was i.i.d. sampled at random from a Gaussian distribution with an unknown grand mean θ and standard deviation ω. To complete the hierarchical model and ensure data-driven posterior inference, noninformative Uniform prior distributions were assigned to all the hyperparameters ($\sigma, \beta, \tau, \theta, \omega$) based on recommendations by Gelman et al. [3].

2.3. Model pooling

Given a belief of exchangeability among the district group-level intercept parameters (4), the hierarchical model allows districts with less information to borrow strength from groups with more information through their shared parent distribution, $N(\theta,\omega^2)$, hereby presenting a compromise between two alternative models: a "no pooling" model and a "complete pooling" model. The no pooling model is the limiting case where the between-group variance parameter $\omega = \infty$, i.e. asserting that there is no information hidden in the between-group distribution of α_k, hereby abandoning the hierarchical modeling of α_k and reducing the model to (1)-(3), generating a separate fit for each district. The complete pooling model represents the opposite limiting case where $\omega = 0$, which arises when separation in the district-level is believed to be irrelevant, imposing the restriction that $\alpha_k = \theta \forall k$, i.e. all districts share the same intercept eliminating the effect of the districts in explaining differences in energy use. All three models – the no pooling model, the hierarchical model and the complete pooling model – were fitted to investigate whether information was hidden in the location of the buildings (effect of district grouping).

2.4. MCMC algorithm for posterior inference

The multi-dimensional joint posterior distribution cannot be obtained analytically; hence, a numerical approach was employed based on Hamiltonian Monte Carlo (HMC), a hybrid Markov Chain Monte Carlo (MCMC) algorithm whose equilibrium distribution is indeed an approximation of the joint posterior distribution [4]. Four chains were run in parallel with randomly dispersed starting points in the parameter space to draw samples from the posterior

distribution. For each chain, 2,000 MCMC samples were drawn with the first 1,000 samples being considered cool, meaning that information about the starting point might still prevail. Samples from this cold period were thus discarded leaving only the warm part of the chains for analysis.

Convergence in the warm chains was monitored in terms of the potential scale reduction factor, $\hat{R}$, for which $\hat{R} \in \mathbb{R}|1 < \hat{R} < \infty$. It is an estimate of the scale with which the variations in the inferred parameter distributions might be reduced if the simulations were continued in the limit $n \to \infty \left(\lim_{n\to\infty} \hat{R} \to 1 \right)$ [4]. $\hat{R}$ accounts for the within-chain and between-chain variance in the warm chains, simultaneously evaluating both the mixing and stationarity of it. For $\hat{R} < 1.1$, a stable and converged estimation was considered for each parameter, respectively.

3. Results

3.1. Model selection

The ability of the three models to fit the data was assessed by means of their expected predictive accuracy in terms of the Watanabe-Akaike information criterion (WAIC), a state-of-the-art fully Bayesian measure of model fit [5]. Compared to non-Bayesian measures like the AIC [6] and BIC [7], and the somewhat Bayesian measure DIC [8], WAIC has the desirable property of averaging over the posterior distributions rather than conditioning on point estimates, making WAIC a fully Bayesian approach for estimating the out-of-sample expectation [9]. As might be expected, the hierarchical model shows to have the lowest WAIC value and thus constitutes the best fit to data (Table 1).

Table 1. Predictive accuracy of the three fitted models.
Lower values of WAIC imply higher predictive accuracy.

	No pooling ($\omega = \infty$)	Hierarchical (ω *estimated*)	Complete pooling ($\omega = 0$)
WAIC	-487	-495	-486

3.2. Explained variance by groups

The hierarchical structure allowed the unknown data variance to be decomposed and fitted to the standard deviation hyperparameters at the three levels of the hierarchical model (σ, τ, ω). The proportion of variance explained at each level of the model was assessed in terms of the intraclass correlation coefficient (ICC), here shown for the first level (variation between annual measurements of the same building):

$$ICC_{level1} = \frac{\sigma^2}{\sigma^2+\tau^2+\omega^2}. \quad (5)$$

The ICC measures (Table 2) indicate that approx. 30%-65% (95% central posterior probability) of the overall data variance can be explained by phenomena between city districts (level three). Adding information about which particular building within a given district the data originates from can explain an additional 25%-51% of the data variance (level two). The remaining data variance is explained by phenomena between the annual measurements of the buildings (level one).

Table 2. Intraclass correlation coefficient (ICC) of groups.

ICC	Posterior quantiles				
	2.5%	25%	Median	75%	97.5%
Within buildings/between years (level one)	0.10	0.13	0.15	0.17	0.20
Within districts/between buildings (level two)	0.25	0.34	0.38	0.43	0.51
Within dataset/between districts (level three)	0.30	0.41	0.47	0.53	0.65

A one-way analysis of variance (ANOVA) was used to test the null-hypothesis that all the district mean energy use distributions, φ_k, were equal (H_0: $\varphi_1 = \varphi_2 = \cdots = \varphi_L$) against the alternative hypothesis that at least one of the districts had a different mean energy use (H_A: $\varphi_1 \neq \varphi_2 = \cdots = \varphi_L$). The null-hypothesis was rejected ($p < 1e\text{-}10$)

indicating the presence of a significant effect of the district location on the annual district heating energy use of buildings.

A model prediction of each district group is shown in Figure 1 using the 37 individual district sub-models (3) of the hierarchical model with individual intercept parameters, α_k. Ignoring the differences in the district groups ($\omega = 0$), the grand mean model is overlaid using the mean intercept, θ.

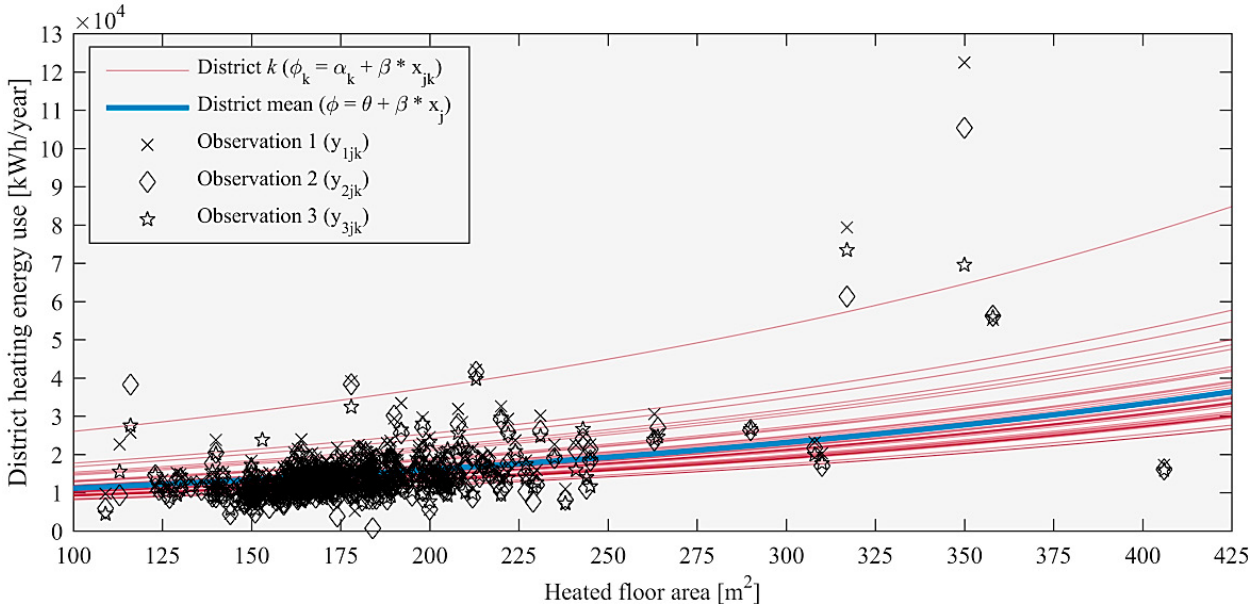

Figure 1. Predictions using the 37 individual district sub-models and the overall mean model.

4. Discussion

4.1. Variability in energy use

The results indicate that energy consuming mechanisms and phenomena between district locations account for 30%-65% (95% central posterior probability) of the overall data variance. This suggests the presence of an additional socio-economic effect causing buildings from different city areas to behave and, to some extent, consume energy differently. Similar conclusions are drawn in another study of the Danish residential building stock [10], where considerable variation in how people consume heat is attributed to various socio-cultural factors. It would be interesting, in future work, to include information about the household income of the different districts to test whether there is indeed a socio-economic correlation between energy use and income. However, district-level variations could also be caused by local regulations on the architecture, environmental exposure, e.g. local wind patterns, shading, etc., and/or because buildings in certain areas are systematically designed as low-energy construction.

The fact that an additional 25%-51% (95% central posterior probability) of the data variance is capsulated and explained by phenomena between buildings within the individual districts is particularly interesting. Bearing in mind that the investigated buildings were all detached single-family houses constructed during a period of unchanged building regulations, such phenomena can hardly be ascribed to any major building-physical or geometrical differences. Instead, it seems legitimate to expect this variation to be caused mainly by the very stochastic nature of occupant behavior. However, to more accurately account for the influence of occupant behavior, additional information about the buildings would have to be included in the model, e.g. number of occupants living in each building, presence of basements and heated attics, energy label ratings, etc.

The remaining 10%-20% (95% central posterior probability) data variance is attributed to energy consuming mechanisms and phenomena between annual measurements within individual buildings. This annual variation is

probably caused by natural variations in the weather. However, strictly speaking, this number contains any residual variability that cannot be decomposed with the proposed hierarchical structure, e.g. observation error.

4.2. Perspectives on future model expansion

Expanding the hierarchical model to include also other building typologies and vintages, e.g. terraced houses and apartment blocks, would allow for a more exhaustive investigation of the mechanisms behind energy use, and for additional research questions to be asked and answered. Bearing in mind that the hierarchical structure allows information to be borrowed from the different sub-groups, such an extended model could potentially provide a more precise and profound estimate of common effects, e.g. occupant behavior. From a more practical point of view, an expanded model could serve as a city-scale prediction tool for e.g. urban planners and municipality managers and would be beneficial for e.g. assessing the capacity of existing public supply systems when expanding building areas within the city, and for assessing the impact of different retrofit scenarios.

5. Conclusion

Significant information about annual heating energy use is contained in knowing the location within a city from which the measurements are obtained. It is proposed that this phenomenon is caused by a socio-economic effect. The effect is not, however, equally profound for all investigated city districts as the energy use in some districts is more similar than in others. In addition, what seems to be a natural variation in occupant behavior between typologically identical buildings explains the majority of variation within city districts, leaving only a minor variance proportion to be explained by weather phenomena.

Acknowledgements

The research was conducted as part of the “Resource Efficient Cities Implementing Advanced Smart City Solutions” (READY) project, work package 3, financed by the 7th EU Framework Programme (FP7-Energy project reference: 609127). Furthermore, the authors would like to thank the district heating company in Aarhus, AffaldVarme Aarhus, for supplying the building energy data that forms the basis of the study.

References

[1] M. H. Kristensen, Danskernes fjernvarmeforbrug [in Danish], Aarhus University, 12 12 2016. [Online]. Available: https://www.youtube.com/watch?v=Opd_hJMQBos&feature=youtu.be. [Accessed 28 01 2017].
[2] R. V. Andersen, J. Toftum, K. K. Andersen and B. W. Olesen, Survey of occupant behaviour and control of indoor environment in Danish dwellings, Energy and Buildings 41 (2009) 11–16.
[3] A. Gelman, Prior distributions for variance parameters in hierarchical models, Bayesian Analysis 3 (1) (2006) 515-534.
[4] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari and D. B. Rubin, Bayesian Data Analysis, 3 ed., CRC Press, 2014.
[5] S. Watanabe, Algebraic Geometry and Statistical Learning Theory, Cambridge, UK, Cambridge University Press, 2009.
[6] H. Akaike, Information theory and an extension of the maximum likelihood principle, in Proceedings of the Second International Symposium on Information Theory, Budapest, 1973.
[7] G. Schwarz, Estimating the dimension of a model, The Annals of Statistics 6 (2) (1978) 461-464.
[8] D. J. Spiegelhalter, N. G. Best, B. P. Carlin and A. van der Linde, Bayesian measures of model complexity and fit, Journal of the Royal Statistical Society. Series B (Statistical Methodology) 64 (4) (2002) 583-639.
[9] A. Gelman, J. Hwang and A. Vehtari, Understanding predictive information criteria for Bayesian models, Statistics and Computing 24 (6) (2014) 997-1016.
[10] A. R. Hansen, The social structure of heat consumption in Denmark: New interpretations from quantitative analysis, Energy Research & Social Science 11 (2016) 109-118.