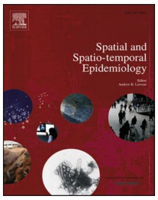
Spatial and Spatio-temporal Epidemiology 43 (2022) 100547

Contents lists available at ScienceDirect

Spatial and Spatio-temporal Epidemiology

journal homepage: www.elsevier.com/locate/sste



Original Research

Bayesian and network models with covariate effects for predicting heating energy demand

Pablo Juan[a,b], Marta Braulio-Gonzalo[c], Carlos Díaz-Ávalos[d], María D. Bovea[c], Laura Serra[b,e,*]

[a] IMAC, Department of Mathematics, Universitat Jaume I, Castellón, Spain
[b] Research Group on Statistics, Econometrics and Health (GRECS), University of Girona, Girona, Spain
[c] Department of Mechanical Engineering and Construction, Universitat Jaume I, Castellón, Spain
[d] Department of Probability and Statistics, Universidad Nacional Autónoma de México, México City, Mexico
[e] CIBERESP, Madrid, Spain

ARTICLE INFO

Keywords:
Covariates
Heating energy demand
INLA
Networks
SPDE

ABSTRACT

The spatial effect is an element presented in many geostatistical works and it should be incorporated into studies regarding the heating energy demand of residential building stocks. The most common approaches have been made by simple descriptive statistics or using analyses by Markov random fields. In this work, we propose two different methods. First, the Stochastic Partial Differential Equation with the Integrated Nested Laplace Approximation to model the variable heating energy demand in Castellón de la Plana, Spain also considering covariates and the spatial effect. Second, simulated street networks for analysing data. We describe and take advantage of the Bayesian methodology in the modelling process in all the scenarios, including covariates and the possibility of creating a simulated street network with the data for the modelling issue. Our results show that the spatial location of the building is a crucial element to study the heating energy demand using both methodologies.

1. Introduction

Heating energy demand (ED_h) has an impact on prices and is a growing concern about environmental problems and global warming. ED_h was recently incorporated into the energy production agenda (Royston et al., 2018). The study of factors associated to changes in ED_h is complex due to its association with other variables such as building block geometry and spatial distribution of building units in urban areas, characteristics of energy distribution networks, morphology of the urban layout and weather-related variables, amongst others.

Office and residential buildings account for a significant proportion of energy demand and use. Residential energy use, mostly for heating, is an important fraction of the total energy use in Spain and energy efficiency action plans have been in operation in Spain since 2014 (IDAE 2016). For these plans to work properly, it is important to monitor the demand and to predict future trends. ED_h shows spatial variability because the spatial distribution of households and different kind of industries is not completely random. This is mainly because the presence of short scale heterogeneity in building and size. In consequence it is expected that trends in energy demand will also show spatial variability. Therefore, the inclusion of spatial heterogeneity in the modelling process of variables showing spatial variation improves model quality and gives the correct power in statistical tests (Cressie, 1993). This is important because municipalities can use the resulting models and the graphical representation of the resulting maps in planning actions for a better use of energy.

In this work, we present the results of the spatial analysis and modelling of the ED_h in the residential building zone of Castellón de la Plana, Spain, and its association to urban and building characteristics (Braulio-Gonzalo et al., 2016). We selected the city of Castellón de la Plana, Spain, because it is a medium-sized city, and it has been chosen before to implement a bottom-up-based model to predict ED_h.

We consider models in the class of the Generalized Linear Mixed Models (GLMM), a class that became popular in the late 80′s and early 90′s to analyse and predict different kinds of response variables with linear association to random and fixed factors (Breslow and Clayton, 1993). In our study we use information on several covariates related to urban and building characteristics at the building level. Our study only covers the wintertime, when the need for heat increases EDh and most of the buildings' global energy use (Eurostat 2018).

[*] Corresponding author.
E-mail address: laura.serra@udg.edu (L. Serra).

https://doi.org/10.1016/j.sste.2022.100547
Received 20 May 2022; Received in revised form 14 September 2022; Accepted 9 November 2022
Available online 11 November 2022

Even if there are other approximations such as Markov Change Monte Carlo (MCMC) or Generalized Linear Models (GLM) (Braulio-Gonzalo et al., 2016), in this study we used the Integrated Nested Laplace Approximation (INLA) approach to identify the variables that affect the environmental performance of the life cycle of Electrical and Electronic Equipment (EEE) as it gives as additional computational advantages.

In our model setting we used the common linear relation between the link function and the covariates. We included a spatially correlated random effect to account for the spatial correlation and clustering commonly observed in spatial data. The random effect is assumed to be a random field with a fixed covariance structure. We compared the usual geostatistical approach in which the random field Z is defined for every point inside the domain D that is the study area, with a simulated network approach, in which Z is defined only at the points of a simulated linear network of city streets. In the traditional geostatistical setting the distance between two locations is the Euclidean distance, whilst for a network, the distance is defined as the shortest path between those locations along the network (Okabe and Sugihara, 2012; Baddeley et al., 2015). Using the simulated linear network approach makes sense because electricity is delivered along a network of electric lines running along the same street network defined for buildings and industries.

The rest of the paper is organised as follows: Section 2 describes the data set used to model the energy demand; Section 3 describes the statistical methodology and provides the details needed to clarify the Bayesian modelling methodology we used. Finally, Section 4 presents the results and discussion.

2. Data set and modelling process

The energy demand data used in this work come from an urban neighbourhood in the municipality of Castellón de la Plana, a medium-sized city with 169,498 inhabitants (INE 2018) located on the east coast of Spain, and an area with mild climate, temperate winters, and warm summers.

The data for this project were obtained from a set of randomly selected buildings in different neighbourhoods of Castellón. ED_h data were obtained by modelling energy efficiency using building simulation software EnergyPlus (U. S. Department of Energy 2015) with the DesignBuilder interface (DesignBuilder UK. 2015). EnergyPlus uses computer-based simulation tools to perform detailed analysis of a building's energy use. These are a commonly used software for conducting energy modelling studies, such as Theodoridou et al. (2011), Caputo et al. (2013), Ascione et al. (2013), Mauro et al. (2015) and Fonseca and Schuleter (2015), amongst others. Herein, the simulations were conducted according to the procedures set in EN ISO 6946:2012 (CEN, 2012) and in EN ISO 673:2011 (CEN, 2011). In this modelling process after doing a comprehensive analysis of the neighbourhood, five covariates were identified to characterise buildings and their urban surroundings. At the building scale the available information was shape factor (S/V) and the building's year of construction (Y). At the urban scale, although there is a variety of covariates that can be used to describe the urban taxonomy, for the city of Castellón de la Plana we only had available spatial information about urban block type (UB), street height-width ratio (H/W) and solar orientation of the main façade (O). The definition of all the covariates considered in this study are presented in Table 1 and were described in detail by Braulio-Gonzalo et al. (2016).

Fig. 1 shows the spatial location of the 574 buildings in Castellón and the histogram of the ED_h data resulting from the modelling process described above. The number of building units per area is higher in the central parts of the city and decreases in the outer zones. The marginal distribution of the ED_h values is skewed to the left, indicating the presence of a high number of buildings with low energy efficiency. Those low ED_h values correspond to old buildings, which lack insulation systems in walls and windows. Old buildings are scattered mostly in the central part of Castellón, as well as in the area close to the port, known as "El Grau" in the eastern part of the city.

Table 1
Covariates of the model.

Scale	Covariate	Characteristics	Description
Building	Year of construction (Y)	• Before 1940	Absence of thermal insulation One layer-thick walls with thermal inertia
		• 1940 – 1959	Absence of thermal insulation Two-layer light walls
		• 1960 – 1979	Absence of thermal insulation Two-layer light walls
		• 1980 - 2006	Poor thermal insulation Two-layer light walls
		• After 2006	Poor thermal insulation Two-layer light walls
	Shape factor (S/V)	• $MF_{T(\leq 4)}$	Multifamily terraced building $\leq$ 4 floors
		• $MF_{T(>4)}$	Multifamily terraced building $>$ 4 floors
		• $SF_{T(\leq 4)}$	Single family terraced building $\leq$ 4 floors
Urban	Solar orientation (O)	• North	Indirect solar radiation
		• East	Direct solar radiation in the mornings; small elevation angle
		• South	Direct solar radiation at noon; maximum elevation angle
		• West	Direct solar radiation in the afternoons; small elevation angle
	Street height-width ratio (H/W)	• H_u=24 m; W_U=10m	Narrow streets that imply poor solar access
		• H_u=24 m; W_U=20m	Wide streets that imply good solar access
	Urban block (UB)	• UB_1	Big internal courtyard that allows solar gains on the south, east and west façades of buildings with an inward orientation towards the courtyard
		• UB_2	No big courtyard, but smaller own light wells as internal building elements

After defining the covariates characterising urban taxonomy, ED_h data were obtained according to the following information:

- **Empirical energy performance assessment**: a sample of three buildings was selected such that the three building typologies in the neighbourhood under study in terms of the covariate Shape factor(S/V) were represented: multifamily terraced buildings with four floors or fewer (MF≤4), with more than four floors (MF>4) and single-family terraced buildings with four floors or fewer (SF≤4. So, the covariates of these three building typologies were identified as follows. The year of construction (Y) was divided in five time periods, because different construction periods imply different thermal transmittance values for the facades, roofs, and floors of buildings. Variations in solar orientation (O) are expected to be associated to different natural heating gains, which notably influence the ED_h of buildings. The more solar gains, the less heating use results. The street height-width ratio (H/W) was included to test if narrow streets (high ratio) or wide streets (low ratio) have different solar gain. Finally, two urban block types (UB) were identified in Castellón: UB_1, with a big internal courtyard that allows solar access on the building's façades; UB_2, with smaller own light wells as internal building elements that block solar access. By combining the five covariates (Y, S/V, O, H/W and UB), multiple variations were conducted. As a

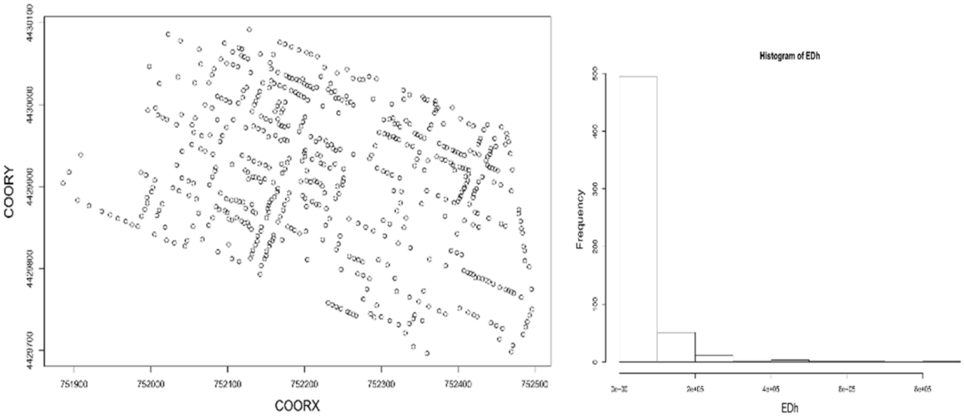


Fig. 1. Data distribution and the histogram.

result, 240 hypotheses were obtained that allowed to extrapolate the conclusions to the rest of the building that integrates the whole neighbourhood under study.

- **Statistical modelling**: The analyses were carried out with the R freeware statistical package (version 3.5) (R Core Team 2016) and the R-INLA package (R-INLA 2017). With the estimates for the response variable, we produced predictions for the ED_h of non-sampled buildings. The applied model is defined in the next section.
- **Stock aggregation**: the individual building prediction results were aggregated to extrapolate conclusions at the urban scale. Here the Geographical Information System (GIS) technology was used, which combines geo-referenced information with cartography, allowing digital maps of urban areas to be developed to identify certain specific aspects of the built environment by a graphical interface. In this study, the cadastral data of the urban area under study was processed by gvSIG software (Asociación, 2022). This means providing the ED_h of each building that comprises the neighbourhood via a graphical scale (Fig. 2), where it can be seen that all possible distances between the measurements have been taken into account, and Energy map for the ED_h of the buildings in the neighbourhood (Fig. 3).

3. Statistical modelling methodology

Generalized Linear Mixed Models (GLMM) were fitted, taking the ED_h data from the sampled buildings as the response variable and a log link function. Besides the covariates included in the study, the model incorporated an independent error term and a spatially correlated error term. The spatially correlated error term was included to capture the spatial variation not accounted by the covariates in the models. The independent error term was included to capture the error induced by the modelling of the ED_h to obtain the data for that variable. Hierarchical Bayesian methods are a common choice to fit GLMM (Blangiardo and Cameletti, 2015 and Braulio-Gonzalo et al., 2016). Instead of using the Markov Change Monte Carlo (MCMC), we use the Integrated Nested Laplace Approximation (INLA) methodology, developed by Rue and Martino (Rue et al., 2009), because it is short computational time and much easier to fit complex models (Braulio-Gonzalo et al., 2016).

The model framework and the process we followed to construct the energy efficiency maps for Castellón is summarised in Fig. 4. Details on the models and the fitting methodology were as follow:

Let Y(s) denote the energy efficiency at location s. The data *{y(s*$_i$*), i = 1,…,n}* were assumed as realisations of a stochastic process indexed by s, this is

$$Y(s) \equiv \{y(s) : s \in D\}$$

where the study area *D* is a subset of R^d.

Unlike ordinary GLMM models (Breslow and Clayton, 1993), where the response variable is assumed independent, the assumption of independence is relaxed through the inclusion of a spatially correlated error term. Spatial data are often associated with other spatial variables or covariates, which might also show spatial variability. The inclusion of the covariates and spatial term permits to account for the effects of risk factors on the spatial distribution of the variable of interest (Aragó et al., 2016 and Serra et al., 2013) and to use the proper power in statistical inferences regarding the model parameters (Cressie, 1993).

In our application, the goal is to model the mean ED_h ($E[Y(s_i)]$) for the building block units, which is related with a linear combination of

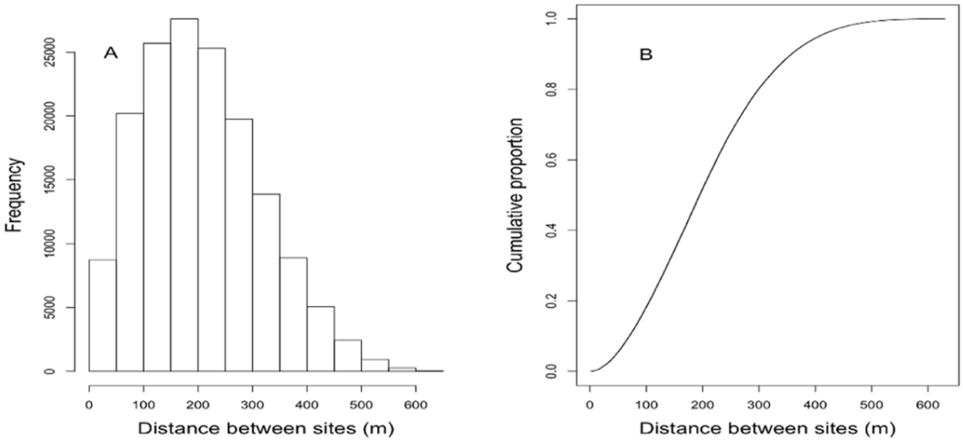


Fig. 2. The histogram of the distances between sites and the cumulative proportion.

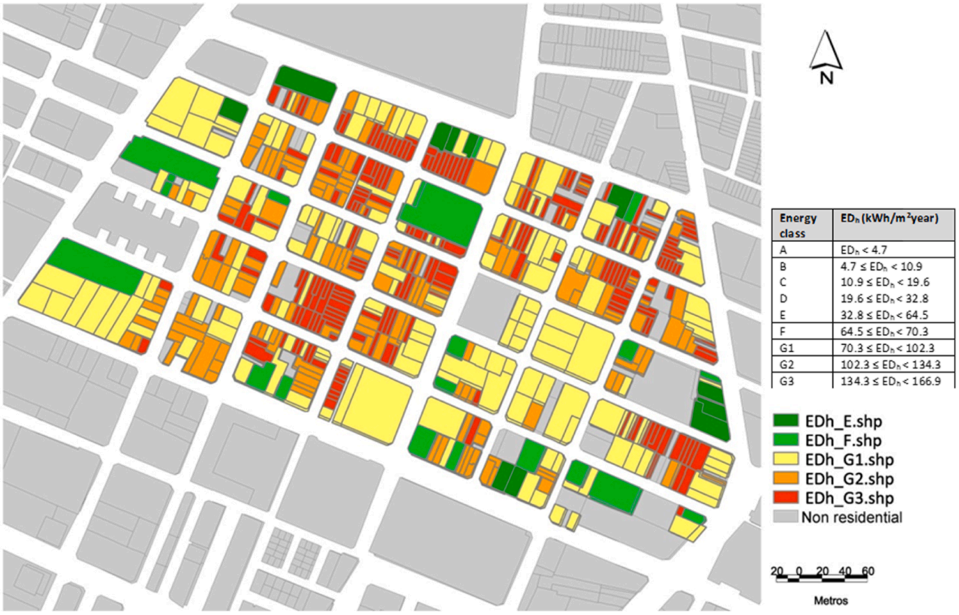


Fig. 3. Energy map for the ED_h of the buildings in the neighbourhood.

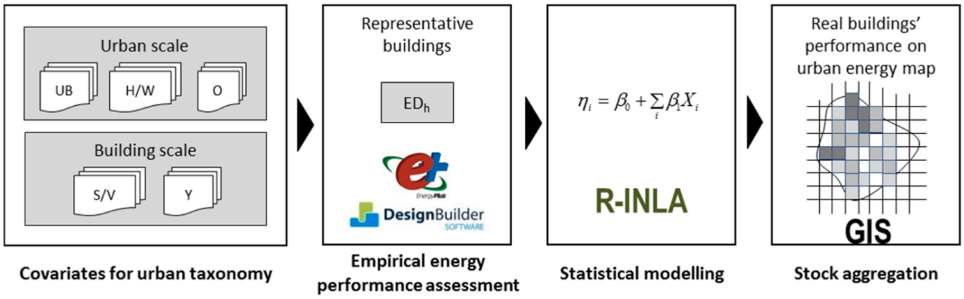


Fig. 4. Model framework and process followed to construct an urban energy map.

the covariates through the relation:

$$\eta_{st} = \beta_0 + \sum_{m=1}^{M} \beta_m z_{m,st} + \sum_{l=1}^{L} f_l(\nu_{l,st})$$

Where $\eta_i = h(\mu)$ is the linear predictor which relates the expected value of the response variable with a linear combination of the covariates and the spatial term through the link function $h(\cdot)$; $\beta = (\beta_0, \beta_1, \ldots, \beta_M)$ are the coefficients that quantify the effect of covariates $z_j = (z_{1j}, .., z_{Mj})$ on the response, and $f = \{f_1(.), .., f_L(.)\}$ is a collection of functions defined in terms of a set of covariates $\nu = (\nu_1, .., \nu_L)$ that include the random effects as well as the spatially correlated effect. This spatial effect is a random effect itself because we don't know its structure completely, and it is introduced as a random field with Matérn covariance structure.

The default value for the smoothness parameter of the Matérn covariance in R-INLA is 2 (Blangiardo and Cameletti 2015). In the same way, the temporal effect (in this case the year) is included as a random walk effect. We used the log link function because the response variable is positive and continuous (Braulio-Gonzalo et al., 2016). We fitted and tested several models using the INLA-SPDE approach (Lindgren et al., 2011).

From this definition, varying the form of the functions $f_l(.)$ we can estimate different kinds of models, from standard and hierarchical regression to spatial and spatiotemporal models (Rue et al., 2009).

3.1. Statistical modelling with INLA-SPDE

The INLA-SPDE approach allows a Gaussian Random Field (GRF) with Matérn covariance structure to be approximated as a discretely indexed spatial random process known as Gaussian Markov Random Field (GMRF) (Lindgren et al., 2011; Serra et al., 2014 and Juan-Verdoy, 2019). GMRF are defined directly by their first- and second-order neighbour structures, and their straight implementation is time-consuming, which leads to the so-called "big n problem". The discrete approximation made when using SPDE offers significant computational advantages over the well-known Markov Chain Monte Carlo methods used to estimate model parameters for GLMM models in the Bayesian context.

The main idea of the SPDE approach consists in defining the continuously indexed Matérn GF X(s) as a discrete indexed GMRF by means of a basis function representation defined on a triangulation of the domain D,

$$S(j) = \sum_{l=1}^{n} \varphi_l(s)\omega_l$$

where n is the total number of vertices in the triangulation, $\{\varphi_l(s)\}$ is the set of basis function and $\{\omega_l\}$ are zero-mean Gaussian distributed weights. The basis functions $\phi_l(s)$ are not random, but rather were chosen to be piecewise linear on each triangle,

$$\varphi_l(s) = \{1 \; at \; vertice \; l \; and \; 0 \; elsewhere\} \qquad (1)$$

The key step is to calculate $\{\omega_l\}$, which reports on the value of the spatial field at each vertex of the triangle. The values inside the triangle will be determined by linear interpolation (Simpson et al., 2016).

Different GLMM models can be obtained depending on the covariate combination considered for each one. The selection of covariates for the different models fitted was done based on their known association with the ED_h as well as on the availability of information at the needed spatial scale for the analysis. Once a battery of competing models is chosen, we used correlation between real data and predicted by the model (near 1), the smaller root-mean square error (RMSE), the deviance information criterion (DIC) and the Watanabe-Akaike information criterion (WAIC) for model selection. The chosen models were those with a low level of complexity and the lowest WAIC and DIC amongst the battery of models compared (Spiegelhalter et al., 2002; Watanabe 2010~~):~~

The process to construct predictive maps of ED_h in Castellón using the SPDE in Euclidean space included a step to choose the best mesh to approximate the random fields that represent the spatial random effects. In our study, the mesh needed for the INLA-SPDE analysis of these data can be produced in three ways: using only the data position, only the boundary, or using both. In this case, and as shown in the next section, the last option was taken because such a mesh resulted in the best model (Fig. 5).

3.2. Network approach methodology

The modelling process described in the previous section considers that the underlying gaussian random field is defined at every point inside the study area *D*. Buildings in the urban areas are arranged along streets which form a network (Fig. 3). Instead of considering that two buildings centred at two locations s_i and s_j are separated by the shortest straight distance between them we will consider that $|s_i$- $s_j|$ is the distance that a pedestrian must walk to go from s_i to s_j. In planar spatial analysis, one is interested in the analysis of spatially varying phenomena within a bounded subset of R^2. Instead, spatial analysis along networks poses a different challenge in the sense that observations occur on a linear network that is a subset of a bounded planar space. Modelling the data using a network approach requires some modifications of the geometrical context, mainly the topology in terms of distances along a street network. The clustering analysis and correlation in such network requires a measure of distance along paths in the network. Common practice is to measure distance by the length of the shortest path in the network (Okabe and Sugihara, 2012).

Formally, a linear network is the union $L = \cup_{i=1}^{N} l_i$ of finitely many line segments in the plane of the form $l_i = [u_i, v_i] = \{w{:}w = t_{ui} + (1 - t)v_i, 0 \leq t \leq 1\}$, where $u_i, v_i \in R^2$ are the endpoints of l_i. Without loss of generality, we assume that for $i \neq j$, the intersection of l_i and l_j is either empty, or is one of the endpoints of l_i or l_j.

A path between locations *u* and u' in L is a sequence $v_0, v_1, \ldots, v_m$, v_{m+1} of the points in the network, with $v_0 = u$ and $v_{m+1} =$ u', so that $[v_i$, $v_{i+1}]$ is a subset of L for each $i = 0,\ldots,m$. The length of this path is

$$\sum_{i=0}^{m} \| v_{i-i+1} \|$$

where $\| \ \|$ denotes Euclidean distance. The shortest-path distance $\delta_{SP}(u,$ u'$)$ between u and u' is the minimum of the lengths of all paths between u and u'. If there are no such paths, which implies that the network is not connected, then δ_{SP} $(u,$ u'$) = \infty$.

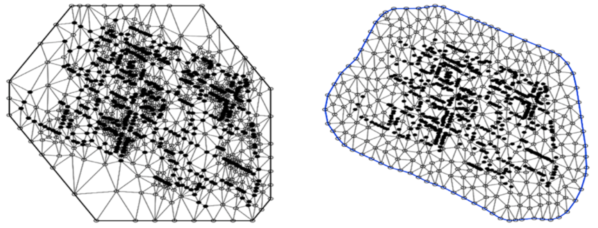


Fig. 5. The two different possibilities of the mesh for the observed data.

In our application, the data for the addresses of the buildings in Castellon were converted to points along a simulated linear network, aiming to resemble Castellon's street network, resulting in a set of points $s_1,\ldots,s_n$ € L, where L denotes the simulated street network of Castellon. Using GIS techniques, we defined a thin buffer zone around L and used this to construct the mesh needed for INLA-SPDE.

Using the buffer zones for the analysis in the Euclidean space and in the linear network space, we fitted several models for ED_h, computing in all cases the goodness of fit statistics described in the next section. Note that the analytical approach we are using is not formally a network analysis, but instead an areal-data approach in a narrow set covering the street network. Nevertheless, we refer to our study area as "network". Models were fitted and tested in a one-by-one basis because there is not an automatic stepwise model selection method implemented for INLA-SPDE.

3.3. Models

For all the models considered in this study, once the mesh for INLA-SPDE was obtained, the next step was to fit different models and compare their diagnostic measures to decide which is the best one under such measures. In all cases, the models were fitted with and without a spatial effect and all included the covariates. The response variable, ED_h, was added in the model as a Gamma likelihood because this error structure was used in previous works (Braulio-Gonzalo et al., 2016). The tested models were:

Model 1 includes covariates (H/W, Y, O, S/V), spatial effect (using model *SPDE*) and temporal effect (using random walk model for year of construction).

Model 2 includes covariates (H/W, Y, O, SV) and temporal effect (using random walk model), without spatial effect.

Model 3 only includes spatial effect (using model *SPDE*).

Model 4 includes covariates (H/W, Y, O, SV), spatial effect (using model *SPDE*) and without temporal effect.

4. Results

The study of the exploratory analysis suggests the need to implement models adequate to capture the presence of an uneven spatial distribution of ED_h. In the next points, the analysis and results are shown.

4.1. Analysis in Euclidean space

The results for the models fitted when we consider that the separation between any two buildings in Castellón is the Euclidean distance between them are presented in Table 1. Goodness of fit for each model was assessed using the criterion of the DIC and CPO. The DIC and CPO values for the four models fitted to the data available are presented in Table 2.

According to the Table 2, in terms of the DIC and the CPO criteria for goodness of fit, the best model is Model 1, this is, the model with spatial component and all the covariates. Although this does not imply that all the covariates included in model 1 are significant, the CPO and DIC

Table 2
DIC, CPO, Correlation and RMSE for the battery of the fitted models.

	Model 1	Model 2	Model 3	Model 4
DIC	12,209.36	12,211.03	12,772.97	12,216.64
CPO	10.83272	10.83404	20.79705	10.83927
Cor	0.661319	0.655411	0.974661	0.660814
RMSE	477,748.7	517,774.5	28,288.13	486,056.6

suggest that their inclusion improves the quality of the model. On the other hand, the correlation, and the root-mean-square error (RMSE) criteria favoured model 3. The best correlation (near 1) and lower RMSE, gives us the best result, and in this case is the model 3, the one with only spatial effect.

When we compare between the observed and the predicted values for the four models fitted models show that the best correspondence between both variables are models 1 and 3 (Fig. 6). Models 2 and 4 show a reasonable quality of fit but over predict large observed ED$_h$ values. This explains in part why they show low correlation and high RMSE.

Table 3 presents the estimates of the covariate effects and their corresponding 0.95 probability intervals. Except for the effect of the building orientation, all of them are significant, indicating that all the covariates are useful to explain in part the variability observed in the energy demand in Castellón. Non-significance of the effect of facade orientation is probably because in most of the city, the residential buildings are over 8 floors tall, and the streets are narrow, so buildings provide shade to each other in facades facing east or west. Increasing the number of floors increases the energy demand as one might expect, due to the increased number of apartments on each extra floor.

Height to width ratio (H/W) has a positive effect on the energy demand. After exponentiation of the coefficient value of 0.2767 for this covariate, we find that an increase of a unit in the H/W value implies an increase of 14.5 percent on the expected energy demand. Regarding the effects of year and shape factor (S/V), they both have a negative effect on the energy demand. Exponentiating the coefficients we find that for a one-year change in the construction of the building the energy demand decreases by 12.5%, this is, buildings constructed in 2000 have an energy demand 12.5% lower than those built in 1999 for instance.

In Fig. 7 we present the maps with the posterior mean estimates of the latent Gaussian random field incorporated in the models (left) and the posterior standard deviation (right). We show only the maps corresponding to models 1, 3 and 4 as the second model did not include a spatial effect. The geographic pattern for the spatial effect with model 1 and 4 looks similar in terms of valleys and hills, but the magnitude of the spatial effect is far lower with model 1, indicating that given the covariates, model 1 gives a better fit to the ED$_h$ data. For model 3, the standard deviation of the posterior estimates of spatial effect is lower near the data points, indicating a poor fitting of such model. This result is consistent with models that do not include external information in the form of covariates, and thus the lowest standard deviation values correspond to the zone where the data were observed, like an ordinary kriging analysis. Models 1 and 4 on the other hand show ups and downs in the study area, as it would be expected.

The scale of the three maps indicates that the lowest standard deviation values correspond to the full model (model 1) as it should be.

Fig. 8 presents the general temporal effect for model 1, showing a clear increase of expected energy demand over time. The energy demand increased strongly during the late 40′s and the 50′s of the past century and remained constant during the 1960′s and even decreasing during the late 1970′s. From there on, the overall trend in energy demand has returned to the values it had at the beginning of the 20th century. It is also clear that energy demand did not increase during the civil war and second world war years.

The parameters related to the random terms for the models (Heterogeneity, Spatial effect, and Temporal effect) are presented in Table 4, as well as their associated posterior standard errors. The temporal term was non-significant for the models that included such component. For the models that included the spatial effect, it was significant using an equivalent 0.05 significance level for its posterior distributions in the different models. Also, the heterogeneity term was significant in all the models including it as a component. The statistical significance of the spatial term indicates that not all the spatial variability is explained by the spatial distribution of the covariates used in the models and that perhaps some other external covariates showing spatial variations should be included in the model. The significance of the heterogeneity

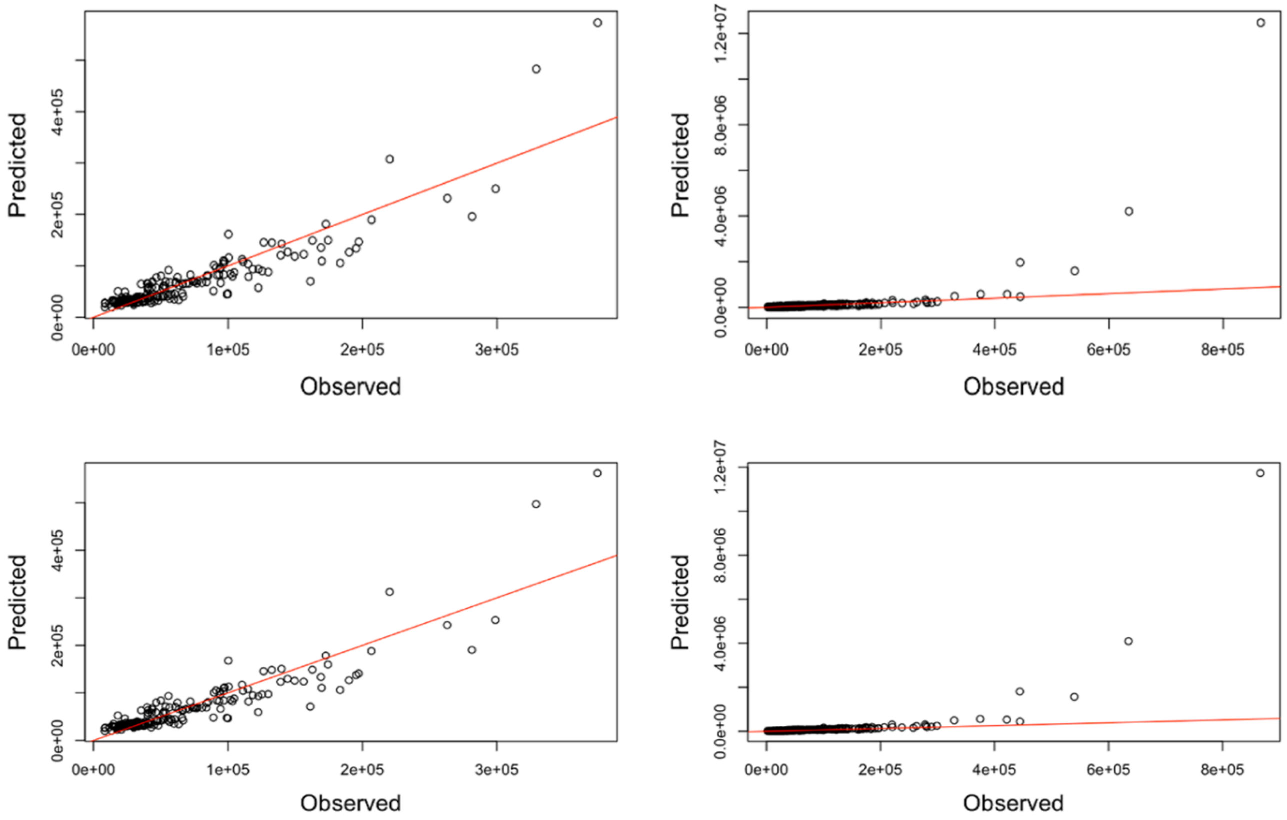


Fig. 6. The Observed vs Predicted ED$_h$ values obtained with the different models. The first line: models 1 and 2. Second line: models 3 and 4.

Table 3
Fixed effects: (mean [0.025quant, 0.975quant]).

	Model 1	Model 2	Model 3	Model 4
Intercept	11.04098 [10.50243,11.60139]	10.99091 [10.47037,11.54501]	10.88097 [10.64596,11.12833]	10.83607 [10.37502,11.29992]
Surface	0.00045 [0.00039,0.00051]	0.00046 [0.00039,0.00052]	–	0.00045 [0.00038,0.00051]
Floors	0.13559 [0.10232,0.16895]	0.13576 [0.10282,0.16786]	–	0.14153 [0.10874,0.17438]
H/W	0.26748 [0.16081,0.37131]	0.27467 [0.17346,0.37396]	–	0.27684 [0.16750,0.38262]
Y	−0.13291 [−0.2445,−0.04164]	−0.12015 [−0.23533,−0.03712]	–	−0.05946 [−0.09521,−0.02353]
O	0.00017 [−0.00017,0.00051]	0.00015 [−0.00019,0.00048]	–	0.00019 [−0.00015,0.00054]

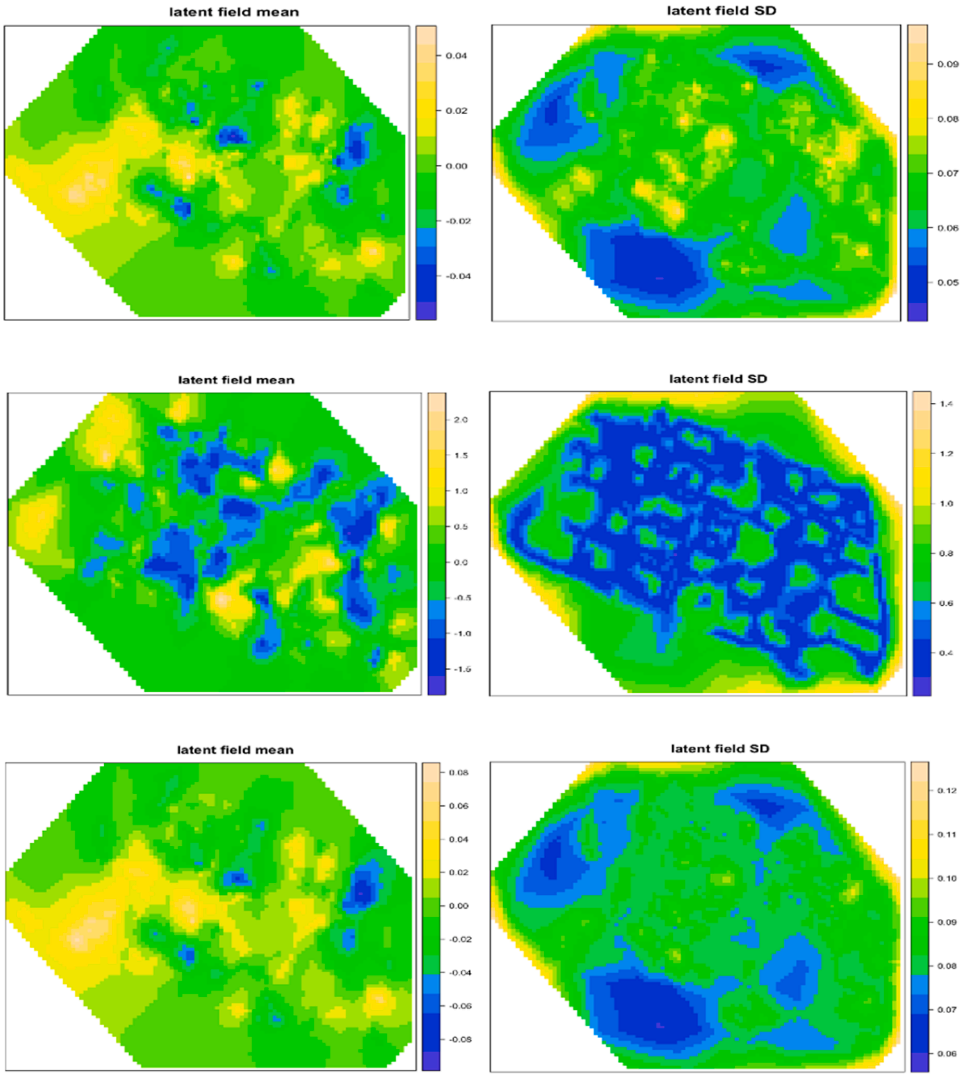


Fig. 7. ED_h prediction maps for the study area obtained with 3 different models. Model 1 (top), Model 3 (middle) and Model 4 (down).

term implies that there is still some variation in the energy demand data that cannot be explained by the model and that it is still subject to further improvement.

In Fig. 9 we present the maps of the ED_h obtained by the different models fitted. Model 3 does not include any covariate in its formulation and in the way the INLA-SPDE algorithm works, it is equivalent to a simple linear interpolation on a grid based on the raw observations. The maps obtained by the rest of the models look similar, but overall, the best fit was obtained by model 1, which includes all the covariates plus the spatial random effect. The highest ED_h values are in areas where building density is low. These areas are associated to places where old single-family farms were located and thus no thermal insulation exist, and farm machinery demands more energy. For the rest of the study area, all the models predict an almost constant ED_h, which is explained by the similar construction techniques used for most of the building and households in those parts of Castellón.

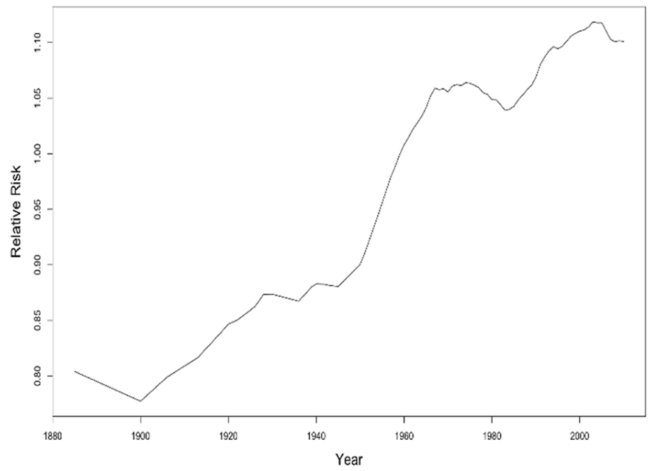


Fig. 8. Estimated temporal effect for the study area showing an increasing trend for ED_h.

4.2. Analysis in network space

Although we don't have the street network for Castellon as a GIS shape file, we simulated linear network with a shape close to the locations of the observed coordinates of the buildings in Castellon. Next, for each building we projected each point to the closest segment of the network. The values for selected covariates for the buildings in Castellón along this simulated network are presented in Fig. 10. Note that the building locations are actual data and only the network of streets has been simulated. Taller buildings tend to be located closer to the edge of the study area, which is the modern part of Castellón and where larger buildings have been constructed in the past 20 years. Regarding energy efficiency typology, the categories are in descending order going from A (highly efficient) to F (least efficient). This covariate is closely related to the presence of means for insulation from external temperature. In Fig. 10 we can see that the least energy-efficient buildings (Class F) are a high proportion of the buildings in our study area and tend to show clusters. Middle-efficiency classes (C and D) are spread regularly over the study area.

To test the significance of the spatial effect, the Berman test is applied (Berman, 1986), whose results indicate that the spatial effect is statistically significant, and it must be included in the models (z1 = 2.7942e-05, p-value=1). A similar conclusion is reached when the spatial Kolmogorov test is used (D = 0.083246, p-value = 0.0008057) (Baddeley et al., 2015). These tests for significance of the spatial effect in the models were significant, indicating that the models must include a spatial effect term. The inclusion of the spatial effect in the models corrects variance estimates and reduces chances of type I error. This confers increased power to statistical tests and allows the correct screening of the external factors that affect energy efficiency.

Finally, a density estimate of ED_h along the network of buildings in Castellón de la Plana is shown in Fig. 11 for two different values of the bandwidth. This estimate is an exploratory tool and provides insight into the non-uniform distribution of energy efficiency. The highest values are observed in the north and west parts of the city, where new housing developments with modern building construction techniques are found. The central and south-eastern parts of the city obtain the lowest efficiency values because these areas are occupied with older buildings and farms, where thermal insulation is poor.

5. Discussion

It is well known that spatial models with covariates have a complexity that makes their fitting non-trivial. Bayesian methods to fit such models have been a useful tool since their appearance in the early 90′s (Besag et al., 1991; Handcock and Stein, 1993). However, the complexity of the models usually forced the use of MCMC methods to obtain samples from the full posterior distributions of the parameters of interest. The INLA-SPDE approach is a faster method to obtain the estimates of the model parameters through the assumption of an underlying gaussian Markov random field that can be estimated approximately using Delaunay tessellations. In our case, the adaptation of the INLA-SPDE to assess the energy efficiency in Castellón, Spain, gave satisfactory results in terms of the statistics used to assess goodness of fit for the models proposed.

The presence of a spatial effect means that the energy efficiency between nearby buildings tends to be similar. This similarity comes from factors such as a similar number of floors, a similar year of construction and a similar building technique, amongst others. Therefore, that areas inside the urban perimeter of Castellon are opened to urban development during similar time laps.

In turn, the assumption that buildings are spread along a network of streets and avenues for the analysis of energy demand makes it a sensible assumption that permits a better analysis in the sense of distance measurement to assess spatial association amongst spatial units and correlation in model error. The inclusion of covariates related to the characteristics of the building stocks in the analysis of the data using the network approach showed to be an attractive way for analysing data observed at locations along city street networks. Such approach has been used previously in the geostatistical context (Abu Bakarra et al. 2016).

From our analyses we have no way to differentiate which of the two approaches could be more advantageous in terms of facility to incorporate the spatial effect. We have shown that for the two approaches presented in this work, it is not difficult to implement models with both, covariates and spatial effects. The choice of the Euclidean or the Network methodology will depend on the nature of the data, this is, if they can be considered as point in a network or if the data can be taken at any point inside a continuous bounded study area.

The use of Stochastic Partial Model fitting with the Stochastic Partial Differential Equation (SPDE) approach, along with the Integrated Nested Laplace Approximation (INLA) to fit the models posted for ED_h in Castellon de la Plana, Spain, has proven a faster and computationally efficient method. It also allows the fit of more sensitive, but complicated models. This is an advantage of our modelling approach, as energy planners need sometime quick responses to changes in energy demand, so to have a model that fits faster but keeps the precision of parameter estimates is desirable in such cases. The inclusion of an error term modelled as a gaussian random field allows the computation of standard errors for the model parameters and the spatial predictions whenever the gaussian assumption is sensible.

However, a possible weakness of our modelling approach is that linear networks for energy supply are not always available for medium and small sized cities, limiting the applicability of the model. Also, building a buffer around the linear network and implementing the INLA approach requires some specialised GIS knowledge.

Table 4
Parameter estimates for the different models fitted under the assumption of data inside a Euclidean space with the corresponding distance metric.

Term	Model 1(mean, sd)	Model 2 (mean, sd)	Model 3 (mean, sd)	Model 4 (mean, sd)
Heterogeneity	(7.3722, 0.4914)	(7.1569,0.4266)	(4.3593, 0.6370)	(7.2572, 0.5161)
Spatial effect	(2.42546, 1.08346)	–	(0.51386,0.13364)	(2.30149, 0.93147)
Temporal effect	(14,213.40, 47,582.98)	(7578.30, 16,029.09)	–	–

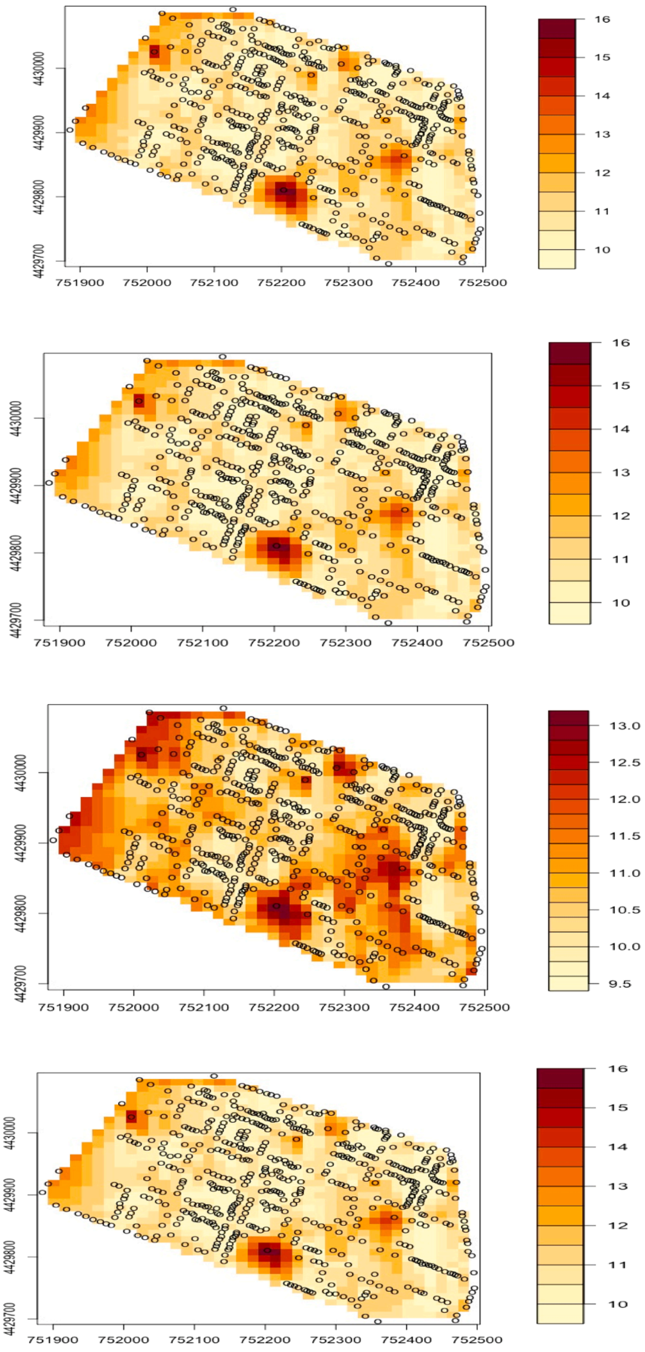


Fig. 9. Prediction of ED_h from model 1 (Top) to model 4 (Down).

Another point is that models fitted under the assumption of the data locations belonging to a Euclidean space and under the Linear network approach are not directly comparable because the data lie on different metric spaces. However, in both cases the best models were those that included covariates and a spatial term, meaning that a substantial part of the spatial variability in ED_h is not accounted by the covariates.

ED_h is related to the demand for electricity generation. Thus, a detailed analysis of ED_h is expected to shed some light on factors contributing to climate change and global warming. Although they are in principle more complicated to fit, models fitted under the linear network approach are more sensible as electric power is delivered through power lines to the buildings and it is such amount of energy that is of interest for ED_h studies. Therefore, for initial studies models for ED_h fitted under the Euclidean approach have the advantage of being easier to fit, but for more detailed and geographically precise studies it is better to use a linear network approach. This will require availability of a GIS shape file of the street network of the city being studied, but such maps are available for most middle size and large cities around the world.

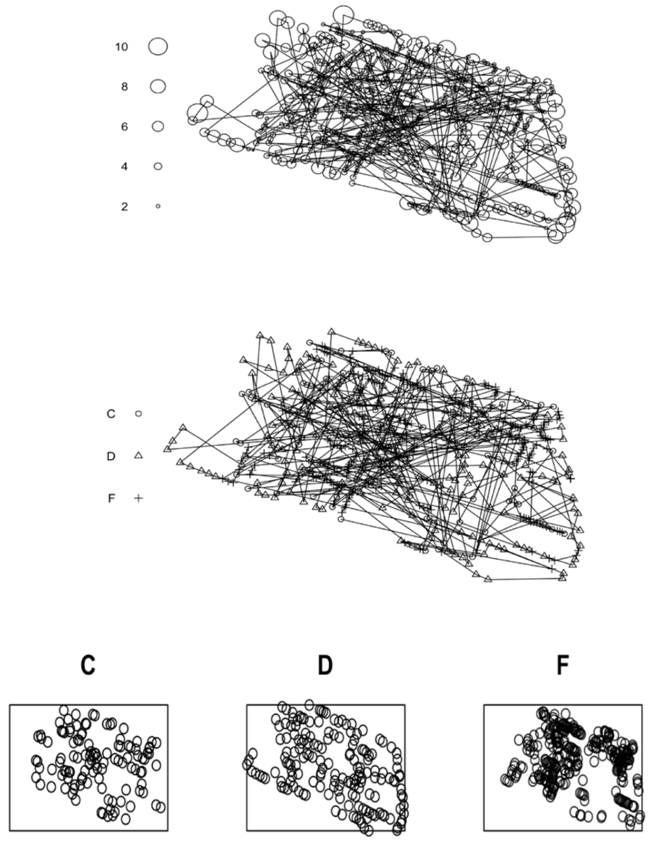


Fig. 10. The building locations and covariate values. The top plot shows the location of buildings in the network number of floors. The middle plot shows the location of buildings by different typology of energy demand. The lower panel shows the different energy demand typology separately.

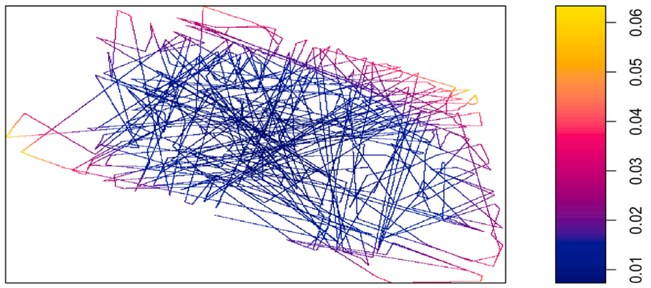


Fig. 11. A kernel estimate of ED_h across the network.

6. Conclusions

The study of the ED_h of residential building stocks is presented herein and indicates the benefits of model estimation and hypothesis testing form including the spatial autocorrelation presented in the geographic data. Previous works have provided simple descriptive statistics. In our case, the methods applied in the analysis enabled ED_h maps to be built, which is very useful material for screening out the covariate effects on the ED_h in the study area.

The assumption that buildings lie along a network of streets to analyse ED_h data is sensible in methodological terms, but a standard methodology has not yet been developed for such kind of data. However, a simulated network analysis in the geostatistical context has been used (Abu Bakarra et al. 2016). The possibility of creating a network and separating data according to the levels of the categorical covariates has been shown.

Finally, the advantages of the Bayesian methodology by creating a network with building locations to allow the use of SPDE and INLA in all the modelling scenarios and the use of models with covariates and a random spatial effect has proven to be a sensible approach to analyse the energy demand in an urban area. The most important elements presented in this work were the assumption that data locations are a realization of a stochastic spatial process in a network (Rakshit et al., 2017), the inclusions of covariates and, finally, the inclusion of a spatial effect. Future research should include the quantitative comparison of the network and continuous approaches of spatial data in urban areas whenever the data are suitable for this purpose.

Funding

The authors thank the funding by Grant MTM2016–78917-R from the Spanish Ministry of Science and Education. Carlos Díaz-Avalos was supported by grant PAPIIT-UNAM IG100221.

Data availability statement

Data sharing not applicable to this article as no datasets were generated or analysed during the current study

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Data availability

Data will be made available on request.

References

Abu Bakarra, J., Sasaki, K., Yaguba, J., Abubakarr Karim, B, 2016. Integrating artificial neural networks and geostatistics for optimum 3D geological block modelling in mineral reserve estimation: a case study. Int. J. Min. Sci. Technol. 26 (4), 581–585.

Aragó, P., Juan, P., Díaz-Avalos, C., Salvador, P., 2016. Spatial point process modelling applied to the assessment of risk factors associated with forest wildfires incidence in Castellón, Spain. Eur. J. For. Res. https://doi.org/10.1007/s10342- 016-0945-z.

Baddeley, A., Rubak, E., Turner, R., 2015. Spatial Point Patterns: Methodology and Applications with R. Chapman and Hall/CRC, London.

Berman, M., 1986. Testing for spatial association between a point process and another stochastic process. Appl. Stat. 35, 54–62.

Besag, J., York, J., Mollie, A., 1991. Bayesian image restoration, with two applications in spatial statistics. Ann. Inst. Statist. Math. 43, 1–59.

Blangiardo, M., Cameletti, M., 2015. Spatial and Spatio-temporal Bayesian Models with R-INLA. John Wiley & Sons.

Braulio-Gonzalo, M., Juan, P., Bovea, M.D., Ruá, M.J., 2016. Modelling energy efficiency performance of residential building stocks based on Bayesian statistical inference. Environ. Model. Softw. 83, 198–211. https://doi.org/10.1016/j.envsoft.2016.05.018.

Breslow, N., Clayton, D.G., 1993. Approximate inference in generalized linear mixed models. J Am Stat Assoc 88 (421), 9–25.

Cressie, N.A.C., 1993. Statistics for Spatial Data. Wiley Series in Probability and Statistics.

Asociación gvSIG. 2022. gvSIG Desktop. Retrieved from http://www.gvsig.com/es/productos/gvsig-desktop.

DesignBuilder UK. (2015). DesignBuilder software.

Eurostat. 2018. Eurostat - statistics explained. Retrieved July 6, 2018, from http://ec.europa.eu/eurostat/statistics-explained/index.php/Energy_consumption_in_households.

Handcock, M., Stein, M., 1993. A Bayesian Analysis of Kriging. Technometrics 35 (4), 403–410. https://doi.org/10.2307/1270273.

IDAE. 2016. Factores de emisión de CO2 y coeficientes de paso a energía primaria de diferentes fuentes de energía final consumidas en el sector de edificios en España. Madrid.

INE. 2018. Spanish Statistical Office. Retrieved from http://www.ine.es/.

Juan-Verdoy, Pablo, 2019. Enhancing the SPDE modeling of spatial point processes with INLA, applied to wildfires. Choosing the best mesh for each database. *Commun. Stat. - Simul. Comput.* https://doi.org/10.1080/03610918.2019.1618473. DOI.

Lindgren, F., Rue, H., Lindstrom, J., 2011. An explicit link between Gaussian fields and Gaussian Markov random fields the SPDE approach. J. Roy. Stat. Soc. Ser. B 423–498.

Okabe, A., Sugihara, K., 2012. Spatial Analysis along Networks: Statistical and Computational Methods. John Wiley & Sons, New York.

R Core Team, 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

R-INLA R-INLA project. 2017. http://www.r-inla.org/home (accessed on November 5th, 2017).

Rakshit, S., Nair, G., Baddeley, A., 2017. Second-order analysis of point patterns on a network using any distance metric. Spat. Stat. 22 (1), 129–154. VolumePart.

Royston, S., Selby, J., Shove, E., 2018. Invisible energy policies: a new agenda for energy demand reduction. Energy Policy 123, 127–135.

Rue, H., Martino, S., Chopin, N., 2009. Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). J. Roy. Stat. Soc. Ser. B 71, 319–392.

Serra, L., Juan, P., Varga, D., Mateu, J., Saez, M., 2013. Spatial pattern modelling of wildfires in Catalonia, Spain 2004-2008. Environ. Model. Softw. 40, 235–244.

Serra, L., Saez, M., Mateu, J., Varga, D., Juan, P., Diaz-Ávalos, C., Rue, H., 2014. Spatio-temporal log Gaussian Cox processes for modelling wildfire occurrence. The case of Catalonia, 1994-2008. Environ. Ecolo. Stat. https://doi.org/10.1007/s10651-013-0267-y. DOI.

Simpson, D., Illian, J.B., Lindgren, F., Sørbye, S.H., Rue, H., 2016. Going off grid: computationally efficient inference for log-Gaussian Cox processes. Biometrika 1 (103), 49–70. VolumeIssue.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van der Linde, A., 2002. Bayesian measures of model complexity and fit (with discussion). J. Roy. Stat. Soc. Ser. B 64 (4), 583–616.

U.S. Department of Energy, 2015. Energy efficiency and renewable energy. EnergyPlus 8.10: energy simulation software. EnergyPLUS. Available from. http://apps1.eere.energy.gov/buildings/energyplus/.

Watanabe, S., 2010. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. J. Mach. Learn. Res. 11, 3571–3594.

CEN, 2012. EN ISO 6946:2012 Building Components and Building Elements - Thermal Resistance and Thermal Transmittance - Calculation Method.

CEN, 2011. EN ISO 673:2011 Glass in Building. Determination of Thermal Transmittance (U Value). Calculation Method.

Theodoridou, I., Papadopoulos, A.M., Hegger, M., 2011. A typological classification of the Greek residential building stock. Energy Build 43, 2779e2787. https://doi.org/10.1016/j.enbuild.2011.06.036.

Cauto, P., Costa, G., Ferrari, S., 2013. A supporting method for defining energy strategies in the building sector at urban scale. Spec. Sect. Long Run Transit. Sustain. Econ. Struct. Eur. Union Beyond 55, 261e270. https://doi.org/10.1016/j.enpol.2012.12.006.

Ascione, F., De Masi, R.F., de Rossi, F., Fistola, R., Sasso, M., Vanoli, G.P., 2013. Analysis and diagnosis of the energy performance of buildings and districts: method- ology, validation and development of urban energy maps. Cities 35, 270e283. https://doi.org/10.1016/j.cities.2013.04.012.

Mauro, G.M., Hamdy, M., Vanoli, G.P., Bianco, N., Hensen, J.L.M., 2015. A new methodology for investigating the cost-optimality of energy retrofitting a building category. Energy Build. 107, 456e478. https://doi.org/10.1016/j.enbuild.2015.08.044.

Fonseca, J.A., Schlueter, A., 2015. Integrated model for characterization of spatio-temporal building energy consumption patterns in neighborhoods and city districts. Appl. Energy 142, 247e265. https://doi.org/10.1016/j.apenergy.2014.12.068.