EasyChair Preprint

№ 1554

Explorative Spatial Data Mining for Energy Technology Adoption and Policy Design Analysis

Fabian Heymann, Filipe Joel Soares, Pablo Duenas and Vladimiro Miranda

September 23, 2019

Explorative Spatial Data Mining for Energy Technology Adoption and Policy Design Analysis

Fabian Heymann[1], Filipe Joel Soares[2], Pablo Duenas[3] and Vladimiro Miranda[1,2]

[1] University of Porto, Faculty of Engineering, Porto, Portugal
[2]INESC TEC, Porto, Portugal
[3]Massachusetts Institute of Technology, 40 Ames Street, 02142 Cambridge (MA), USA
fabian.heymann@fe.up.pt

Abstract. Spatial data mining aims at the discovery of unknown, useful patterns from large spatial datasets. This article presents a thorough analysis of the Portuguese adopters of distributed energy resources using explorative spatial data mining techniques. Results show clustering of distributed energy resources that currently passing the early adoption stage in Portugal. Furthermore, spatial adoption patterns are simulated over a 20 year horizon, analyzing technology concentration changes over time while comparing three different energy policy designs. Outcomes provide useful indication for both electrical network planning and energy policy design.

Keywords: Diffusion of Innovation, Renewable Energy, Spatial Data Mining

1 Introduction

Recently, residential consumers have been adopting new distributed energy resources (DER), energy technologies like solar photovoltaics (PV), electric vehicles (EV) and electric heating, ventilation and air conditioning devices (HVAC). With the constant growth in utilization of these technologies experienced in recent years, studies started to focus on the likely impact of such DER, as they are expected to substantially reshape the European energy system [1]. With its Clean Energy For All Europeans package the European Union (EU) aims at further strengthening the role of consumers and local energy communities within an increasingly decentralized European energy system [2].

Installation and subsequent adequate operation of DER offers several potential benefits, including self-consumption, arbitrage trade, shifted consumption and flexibility provision [2]. The Portuguese government has recently committed to a renewable energy transition as outlined in its National Energy Plan towards 2020 [3]. This strategy includes several goals, i.e. i) to increase the share of renewable energy on final energy consumption by 40% until 2030 and ii) to promote microgeneration, mostly by PV; the target is 300 MW of microgeneration by 2020. Furthermore, ambitious targets to stimulate the uptake of electric mobility technologies is contained.

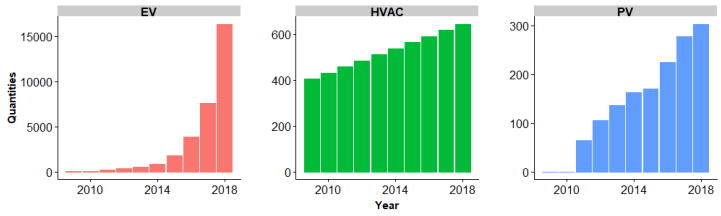

Figure 1. EV, HVAC (both adopters) and PV (in MW) stock in Portugal 2009-2018

Given the political backing, a strong uptake of DER could be observed in Portugal (Figure 1).

This motivates a thorough and combined assessment of the spatial patterns of PV, EV, and HVAC adoption. These patterns are especially interesting to electricity network planners, whose job is to connect consumers to generation sources, thus satisfying the rising electricity needs at a constant, balanced and cost-optimized pace. With regard to energy technologies such as DER, network planners are especially interested in understanding timing and magnitude of the appearance of such new appliances [4]. Therefore, the assessment of spatial DER adoption patterns can contribute with important insights both to policy analysis and electricity network planning [5]. Having the analysis of DER adoption patterns as main topic of this paper, the addressed research questions can be differentiated into the following:

- Does the adoption process of EV, HV and HVAC exhibit spatially clustered or homogeneous patterns?
- Can a decomposition of spatial autocorrelation patterns identify EV, PV or HVAC adoption hotspots (or coldspots) that are statistically significant?
- If we simulate technology adoption over time, how are spatial autocorrelation measures (e.g. Moran's I) evolving?

The report is structured the following way (Figure 2): While Chapter 2 is dedicated to a short introduction of the emerging field of spatial data mining, Chapter 3 includes a description of the data-set analysed. Then, Chapter 4 introduces the mathematical framework of spatial autocorrelation and the detection of localized clusters used throughout this work. Thus, this part of this work is dedicated to what has been called "Exploratory Spatial Data Mining" [6]. Chapter 5 builds on previous explanations and investigates the temporal variability of Moran's I along a 20-year timespan, with a simulated full technology adoption. Chapter 6 contains conclusions and an outlook on the further applications of the developed application of the previous chapter.

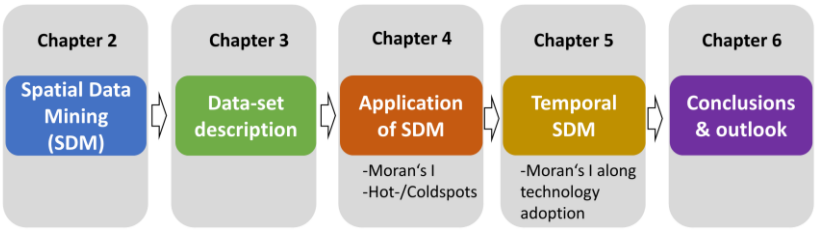

Figure 2. Model architecture

2 Spatial Data Mining

Spatial data mining is an emerging, very recent research area that has developed on top of data mining research which itself exists since the 1980s [7]. Spatial data mining has been defined as the "the process of discovering interesting and previously unknown, but potentially useful patterns from large spatial datasets" [8]. The main difference to traditional data mining thus lies in the very nature of spatial data. Spatial data intrinsically relate to space, therefore carrying information on location, distance and topology/form of spatial objects [7]. Therefore, spatial data is also always assessed with tools that aim to analyze these specific spatial characteristics, such as spatial join, spatial overlay and spatial intersection among others. An extensive overview of the computation of spatial pattern analysis can be found in [9]. In contrast to non-spatial data mining, spatial data mining was said to possess increased complexity [8]. Its main challenges are [7]:

1) Spatial autocorrelation phenomena
2) Spatial relationships between observations and their description
3) The inherent complexity of spatial data

Consequently, the increased complexity and a high computational costs of spatial data processing require the use of efficient spatial data structures and operations [7].

There have been several methods developed that are able to explore and handle the above-mentioned complexities. These can be divided into Exploratory Spatial Data Mining and other advanced techniques.

In [6], the authors merge the common spatial clustering tools such as Global Autocorrelation (Moran's I), Hot Spot (Getis -Ord) Analysis, Local Autocorrelation (e.g. Anselins Local Moran I) and Density kernel estimation as Exploratory Spatial Data Mining analyses. On the other hand, interestingly, more advanced tasks of spatial data mining are the analysis of spatial association rules, spatial clustering analysis, spatial trend detection, and spatial outlier analysis [6].

This work will proceed with a thorough Exploratory Spatial Data Mining of DER adopters in Portugal, followed by the establishment of an advanced tool that allows tracking the variation of spatial autocorrelation structures along a full DER adoption lifetime using a technology diffusion model.

3 Input data

In the scope of this work, two data-sets have been combined. A set of geo-referenced EV, PV and HVAC adopters (counting 2,632/ 474/ 2,111) as point information, obtained by the Portuguese e-mobility charging platform operator and the Portuguese energy agency as well as a highly granular census data-set for Continental Portugal.

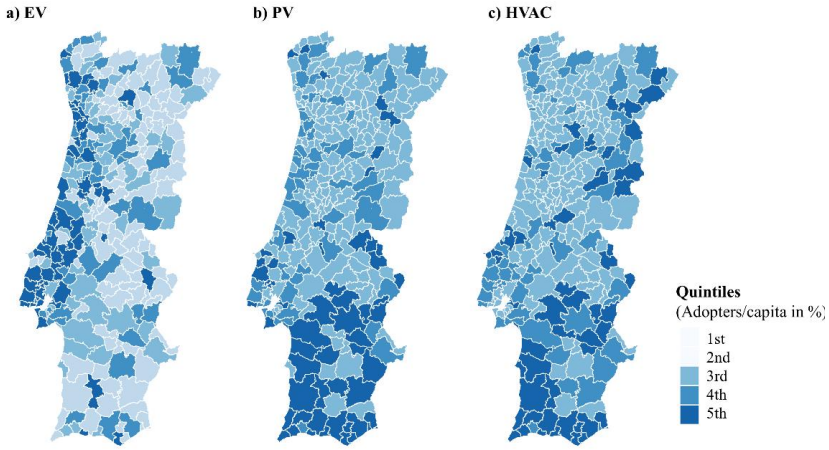

Figure 3. Spatial distribution of EV, PV and HVAC adopters in municipalities of Continental Portugal. Values are divided into quintiles, (for each technology separately).

The EV data-set has been cleaned from commercial users beforehand and sums both Battery EV (BEV) and Plug-In Hybrid EV (PHEV). Likewise, residential PV installation records have been obtained by removing small-scale business. No technology or installation size differentiation have been considered. On the other hand, the HVAC data-set comprises mostly electrical systems of different technical peculiarities. The distribution of selected energy technologies across Portuguese continental municipalities is shown above (Figure 3.)

While this study differentiates the adoption tendencies between this larger EV, PV and HVAC groups, in-group analysis (e.g. BEV versus PHEV) has been constrained by the shape of data-sets available and lies outside the scope of this work.

4 Applying Spatial Data Mining to DER adopter patterns

4.1 Global autocorrelation

A widely-applied metric for spatial auto-correlation is Moran's I [9]. Similar to Geary's C or the global Getis-Ord G, it is an autocorrelation test like that is applied on a global scale Thus, it results in one index. Moran's I provides insight in the observations' tendency having similar (or, linear correlated) values when compared to their neighbors. Input data can be of spatial point or polygon type.

It is an dimensionless, appealing metric, as it produces outputs within [-1,1], where a value of 0 spatial randomness equivalent to possessing no distinct pattern, 1 represents absolute spatial autocorrelation, and -1 complete dissimilarity similar to a checkerboard pattern [10]. A Moran's I value of -1 implies that all spatial objects are neighbored by the most dissimilar values of the population. An important input represents the weight

matrix w_{ij} that contains the neighboring structures of the spatial points or polygons under analysis.

Considering spatial polygons, the neighborhood structure incorporates the degree of adjacency, taking values of 0 (is not neighbor) or 1 (is neighbor). As shown below (Eq.1), the formula sums up all differences between polygons (*i*) values y_i and respective neighborhood polygons' (*j)* values y_j compared to the global mean $\hat{y}$ (also *lagged mean* or *spatial lag*). The resulting value is divided by the variance of each value y_i with respect to the global average $\hat{y}$ and consecutively multiplied with the number of observations *n* by the spatial weights' matrix w_{ij}. Moran's I it is typically computed as stated below:

$$I = \frac{\mathrm{n}}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i-\hat{y})(y_j-\hat{y})}{\sum_{i=1}^{n}(y_i-\hat{y})^2} \quad (1)$$

The output gives a first indication of the spatial autocorrelative structures. However, respective significance levels (*p*-values) can be either obtained through simulation approaches or by comparing the variances to predefined distributions. Former approach has been explained in [11].

It is important to note that Moran's I values can be computed on the base of spatial point data only. However, in our research question such analysis is limited given the adopter's location dependency from population variables, that is, the spatial distribution of population groups that represent the overall adopter potential. Therefore, spatial point information previously introduced has been superimposed with municipality polygons. This approach allows for the retrieval of a DER adopters/1000 inhabitants ratio, which is better suited to compare the presence of EV, PV or HVAC adopters in differently populated areas to each other.

Moran's I values and p-values obtained for the polygon-based analysis are shown below (Table 1). As shown, results suggest that all three technologies exhibit spatial autocorrelation (Moran's I between 0-1). That means that values are similar to neighboring values, or in other words, are spatially clustered.

Table 1. Spatial autocorrelation for EV, PV and HVAC adopters in Continental Portugal

Value/ Technology	EV	PV	HVAC
Moran's I	0.42346	0.37526	0.39532
p-value	<0.01	<0.01	<0.01
Polyg. with values	185	130	161

The p-value has been computed to quantify the probability that the calculated Moran's I values are different from pure chance. As mentioned above, the p-value is approximated using a simulation-based approach firstly presented in [12]. Here, the probability of obtaining Moran's I values above the observed one is calculated using the following formula:

$$p = \frac{m+1}{M+1} \quad (2)$$

In this equation, *m* quantifies the number of simulated Moran's I values above the retrieved value. Furthermore, *M* represents the total number of simulations. As shown in Figure 5, one-sided exceedance probability distributions for obtaining values larger than the retrieved Moran's I have been generated. In other words, the approach simulates a predefined number of Moran's I values relying on observed values in a permutated way. Thus, it can be observed if such value distributions follow a spatial randomization that is equivalent to accepting the Null hypothesis.

The distribution of the exceedance probability was generated using during 600 permutations with equal probability and no repetition. It should be noted that the number of permutations needs to be smaller than the possibilities of rearranging the polygon values to avoid double counting effects that could affect negatively results. This is true for this work.

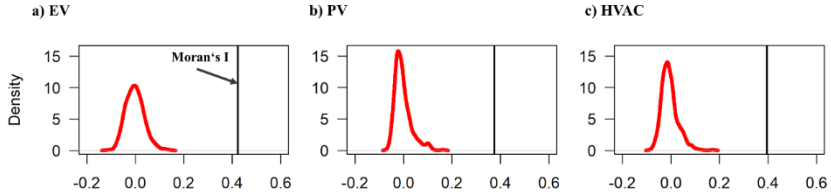

Figure 5. Spatial distribution

As seen above, outcomes suggest strong evidence for the rejection of the Null hypothesis of spatial randomization with p-values smaller than 0.1% across all technologies. The red vertical bar in the figure (Fig.5) indicates the realized Moran's value and the estimated Moran's I value for each technology on Municipal aggregation level. It should be noted that Moran's Is dependency on a predefined neighborhood structure as well as its boundary polygon configuration in an incomplete neighborhood matrix have been criticized in [9]. However, as authors likewise admitted, no optimal treatment of these cases has been found so far.

4.2 Local autocorrelation

While the previously introduced analysis of spatial autocorrelation (Moran's I) provides insight in the global dispersion / concentration of spatial patterns, attempts have been made to break geographical variation down to study local situations.

In this light, Anselin suggested a new type of model, namely the so called "local indicators of spatial association (LISAs)" [13]. These should comply with two requirements:

- The LISA value of each observation should provide insights to the spatial clustering around that value
- The sum of all LISA observations should be proportional to a global metric of spatial autocorrelation (e.g. all LISA values should sum to a global autocorrelation value).

Latter requirement can be met using index decomposition techniques. In the same work [13], Anselin suggested a LISA based on the decomposition of Moran's I, to retrieve a Local Moran's I. Here, the autocorrelation value associated to each observation is I_i, whereas q_i are the mean-centered values and q_j are the means for all neighbor values of polygon i. Thus, I_i can be retrieved following:

$$I_i = q_i \sum_j w_{ij} q_j \tag{3}$$

Using a permutation Monte-Carlo sampling approach as in the test-statistic approach of Eq.2, a significance test may be conducted using [13]:

$$z(I_i) = \frac{I_i - E[I_i]}{\sqrt{Var[I_i]}} \tag{4}$$

Here, values of $I_i > 0$ indicate that a cluster of similar values (higher or lower than average) is present. Likewise, values of $I_i < 0$ indicate a combination of dissimilar values (e.g. high values surrounded by low values). In R programing language, this can be computed using the "localmoran" command of the *spdep* package. This command returns the local Moran's I statistic for each polygon, the expected value $E(I_i)$ and variance $Var(I_i)$ under the randomization hypothesis, the test statistic (Eq. 4) as well as the p-value of the above statistic assuming approximate normal distribution [11].

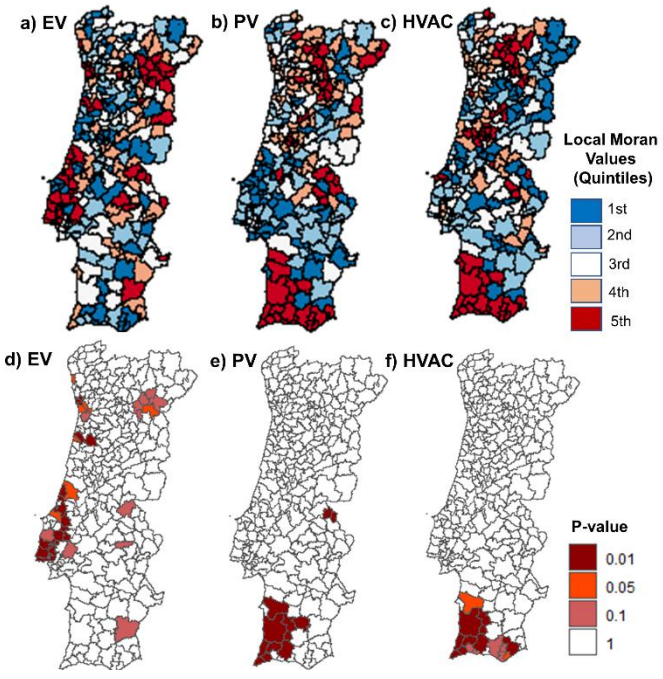

Figure 6. Spatial distribution of Local Moran I (a, b, c) and respective p-values (d, e, f). Please note that a, b and c are shown in quintiles that are equidistant to 0.

Outcomes show spatial hot-spots along the southern cost (PV, HVAC) and Western costs (EV) as well as in some isolated areas in northern-central Portugal (EV, PV, HVAC). Furthermore, all technologies adopter distributions suggest cold spots in the Northern or central areas of Continental Portugal (EV, PV, HVAC). Taking the test statistics analysis into account, the hotspots along the urban centers at Portugal's western coastline (EV) and the southern costal hotspots (PV; HVAC) suggest being significant at levels <1%.

Several techniques to extend the local autocorrelation analysis, taking into account common critiques on the necessary normality assumption (of I_is) and multiple hypothesis testing have been proposed. The interested reader might find an extensive overview of such extensions together with case study applications in [11].

5 Spatial autocorrelation describing technology diffusion

An interesting further application of Moran's I to study large-scale DER diffusion patterns geographically, lies in its potential to describe the stage of maturity of the innovation diffusion process based on analysis of spatial autocorrelation of the adoption patterns. In a thought experiment (Figure 7), we would expect to see isolated adoption clusters at early adoption. Such patterns would possess low spatial autocorrelation. However, autocorrelation would rise (towards higher spatial autocorrelation) during mid-time of the diffusion process, while a mature technology diffusion would likely see equally distributed DER per capita shares.

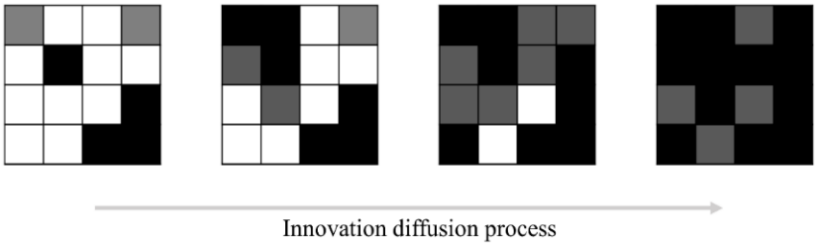

Figure 7. Hypothetical technology diffusion process with spatial patterns of adoption stages. Here, white represents "No adoption", light grey "Intermediate adoption" (33%), dark grey "Strong adoption" (66%), and black "Full adoption" (100%).

Hence, spatial correlation, and, especially its change over time, can be used to study and compare technology diffusion processes across countries or case studies. Furthermore, the strength of autocorrelation (maximum value of Moran's I) and its development over time are interesting inputs to benchmark energy policy designs, in case decision makers are concerned about adoption asymmetries across populations.

Therefore, we establish a temporalization of Moran's I model, which compares 20 annual snapshot DER distributions along the technology diffusion process. While the DER diffusion model has been presented in [5], we further analyze three policy designs that result in different spatial DER distributions. The modelling rationale behind all

Table 2. Modeling of DER distribution under policy change

DER adopter distributions	High weights on the following census variables
High-performance high income (HP)	Census cells with above average higher education share, large residencies, high shares of housing owners
Low-medium income (LMI)	Census cells with below average higher education share, small and old residencies, high shares of renters
Randomized distribution (RN)	-

three policy designs lies in the different ranking of census cells in the queue of DER adoption. While the model uses a global technology forecast for each year as well as spatially granular (neighborhood level) census data, it produces DER shares per municipality for each of the 20 analyzed years. The policy designs are: "high-performance high income (HP)", "Low-medium income (LMI)" and "randomized distribution (RN)". They can be discriminated considering different rankings of census cells that are due next in the adoption process (Table 2). The census variables that have been used to construct the different census cell rankings (through attributing different weights to each household-count normalized census cell) are shown below.

Results are shown below (Figure 7). One can observe spatial autocorrelation growing along the EV adoption process (left side). Interestingly, patterns on the left side (along adoption years) are not linear or follow the expected bell curve outcome (from low adoption – low autocorrelation to medium adoption – high autocorrelation to complete adoption – low autocorrelation). Instead, under a HP policy design, Moran's I increase until year 10 and then fades out to a higher autocorrelation level (0.6). On the other hand, LMI and RN policies see different Moran's I evolutions that stepwise proceed towards a 0.4 or 0.5 autocorrelation value respectively.

Moran's I variations on time are very different if compared to EV adoption shares; as DER adoption is not constant over time, such analysis shows remarkable differences. While under a HP policy design, Moran's I would quickly increase after around 5-10% of adoption, it quickly reaches a plateau. Towards the end of the adoption process, autocorrelation reduces slightly again.

On the other hand, for LMI and RN policies, spatial autocorrelation remains initially on lower levels (around 0.4) and only reach higher levels (such as the HP plateau) after 90% of EV adoption. That is an interesting outcome as it suggests that LMI and HP trigger less autocorrelated (i.e. more dispersed) DER adoption behavior that might eventually reduce system integration costs.

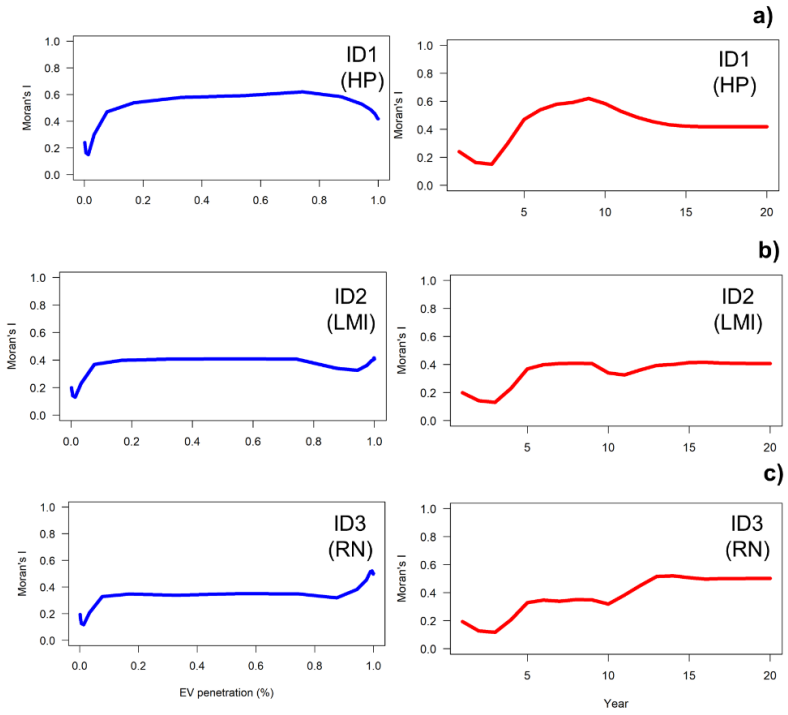

Figure 7. Global Moran's I values for EV adoption patterns along the years (left side) and relative to EV adoption shares (right side). Results show temporal variations for HP, LMI and RN policy designs (a, b, c respectively).

6 Conclusions and outlook

This paper presented a thorough analysis of Portuguese DER adopters using spatial data mining techniques. While adoption patterns on municipal level have been characterized using Moran's I and Local Moran's I, the temporal evolution of spatial autocorrelation behavior using EV as a case study has been analyzed, too.

Results show similar degree of autocorrelation for all three technologies under analysis (EV, PV, HVAC). Likewise, local clusters (<1% significance level) have been detected, with EV hotspots along the western, densely populated coastline and PV and HVAC mostly along the southern fringe along the Algarve coastline.

On the other hand, this paper provided insights in the effects of policy choice on spatial autocorrelation structures. Using a technology diffusion model, outcomes suggested that the choice of energy policy measures has strong effects on spatial autocorrelation structures (and thus inequality and network impact) of DER adopters. For example, policy schemes LMI and RN achieve only higher autocorrelation levels after 10 years of adoption. Thus, such incentive schemes might allow deferring electricity network investments given adoption patterns with a higher degree of dispersion. In contrast, one would expect to associate HP adoption patterns with a more

pronounced investment to integrate EV charging. This is, as such incentive scheme would produce highly concentrated adoption patterns from early adoption phases onwards. Concluding, this study serves as promising starting point to further investigate the dependency among spatial autocorrelation behavior of DER adoption patterns, network expansion costs and adoption inequality.

Acknowledgement

The authors gratefully acknowledge the provision of data-sets by the Portuguese Energy Agency (ADENE) and CEiiA. F. Heymann acknowledges the financial support granted under FCT-MIT Portugal Scholarship PD/BD/114262/2016. This work has been co-financed by National Funds through the Portuguese funding agency (Fundação para a Ciência e a Tecnologia) within project: UID/EEA/50014/2019.

References

[1] K. Bell, "Methods and Tools for Planning the Future Power System: Issues and Priorities," *Model. Requir. GB Power Syst. Resil. Dur. Transit. to Low Carbon Energy*, no. Paper 5, pp. 1–35, 2015.

[2] European Commission, "Clean Energy For All Europeans," Brussels, 2016.

[3] Portuguese Republic, "Plano de Acção Nacional para as Energias Renováveis ao abrigo da directiva 2009/28/CE," 2010.

[4] F. Heymann, J. Melo, P. D. Martínez, F. Soares, and V. Miranda, "On the Emerging Role of Spatial Load Forecasting in Transmission / Distribution Grid Planning," in *11th Mediterranean Conference on Power Generation, Transmission, Distribution and Energy Conversion (MEDPOWER 2018)*, 2018.

[5] F. Heymann, J. Silva, V. Miranda, J. Melo, F. J. Soares, and A. Padilha-Feltrin, "Distribution network planning considering technology diffusion dynamics and spatial net-load behavior," *Int. J. Electr. Power Energy Syst.*, vol. 106, pp. 254–265, 2019.

[6] A. Bhardwaj, "Spatial Data Mining," in *Data Mining Techniques and Tools for Knowledge Discovery in Agricultural Datasets*, IASRI, 2012, pp. 153–166.

[7] K. Koperski, J. Adhikary, and J. Hau, "Spatial Data Mining: Progress and Challenges Survey paper," 1997.

[8] S. Shekhar, P. Zhang, and Y. Huang, "Spatial Data Mining," in *The Data Mining and Knowledge Discovery Handbook*, Springer Publishing Company, Incorporated, 2005, pp. 833–850.

[9] E. Gómez-Rubio, Virgilio; Bivand, Roger; Pebesma, *Applied Spatial Data Analysis with R*, Second Edi. 2013.

[10] M. Gimond, "Intro to GIS and Spatial Analysis [ebook]," 2019. [Online]. Available: https://mgimond.github.io/Spatial/index.html.

[11] L. Comber and C. Brunsdon, *R for Spatial Analysis and Mapping*. SAGE Publications Ltd.

[12] A. C. A. Hope, "A Simplified Monte Carlo Significance Test Procedure," *J. R. Stat. Soc. Ser. B*, vol. 30, no. 3, pp. 582–598, 1968.

[13] L. Anselin, "Local indicators of spatial association - LISA," *Geogr. Anal.*, vol. 27, pp. 93–115, 1995.