Supplementary material

Introduction

This document contains the supplementary material for the paper entitled "Unveiling spatial and temporal patterns of second home dynamics: a Bayesian spatiotemporal analysis for a Mediterranean island". Section A.1 details the definition of spatial data types, and provides tables and graphs for descriptive statistics. Section A.2 provides general definition of the Bayesian hierarchical model and technical details for model selection tools. Section A.3 describes the procedure of the robustness check.

Appendix A.1: Descriptive statistics on Corsican second homes incidence rates

Appendix A.1.1 Spatial data types

Following Cressie (1993), we distinguish two types of economic data: point-referenced data and areal data. The former refers to individual-level data, e.g., housing transaction prices. The latter is based on aggregated-level data, e.g. GDP, unemployment rates. Typically, areal data own two core features: they are built on a given region which is decomposed into many non-overlapped sub-regions, and data are collected through aggregated counts within each sub-region. In our study, we center on areal data, because we are interested in investigating second home rates.

Appendix A.1.2 Tables and graphs

Table A.1. Descriptive statistics for second homes and total houses by county (period: 2007 - 2016).

	Study Region	Dissemination Area			
	Current Total Count	Mean	Min	Max	SD*
Second home	91,622	222.88	1	6,748	461.52

counts					
Total house counts	245,851	619.34	15	33,895	2,132.65
Rate	0.372	0.51	0.022	0.83	0.16

Table A.1 shows second home counts and total house numbers over the study period. Even though the second home count reaches 91622 with the corresponding rate equaled 0.372 at the end of the study period, there is a strong variability within sub-regions (counties). For instance, the minimal and maximum second home rate equals 0.022 and 0.83 respectively.

Table A.2. Descriptive statistics for the temporal variation of second home counts.

	Mean	Min	Max	SD
2006	194.183	1	4474	386.796
2007	197.330	6	4465	389.425
2008	203.175	4	4906	413.219
2009	207.750	4	5077	426.535
2010	212.553	4	5115	432.917
2011	220.594	4	5375	450.401
2012	229.789	4	6025	482.729
2013	236.519	4	6025	486.182
2014	244.769	4	6784	524.319
2015	250.450	4	6539	524.530
2016	254.512	4	6581	529.661

Table A.3. Data source.

Variable	Description	Source
Physical landscape	Physical landscape counts within a county e.g., lakes, mountains, alpine rocks, estuary	Corsica Recreational Areas database, UMR LISA
Cultural landscape	Cultural landscape counts within a county e.g., castles, city walls, towers, churches	Corsica Recreational Areas database, UMR LISA
Coastal county	Dummy variable, 1: coastal county; 0 otherwise	Corsica GIS database, UMR LISA
Mountainous county	Dummy variable, 1: the average elevation ≥ 500m; 0otherwise	Corsica GIS database, UMR LISA

Population	Measured by household	www.insee.fr/fr/information/2008354
Household growth	$(household_{it} - household_{it-1}) / household_{it-1}$	www.insee.fr/fr/information/2008354
Interest rate		webstat.banque-france.fr/en/quickview.do?SERIESKEY=243.MIR1.Q.FR.R.A22FRX.A.R.A.2254FR.EUR.C50
Council tax	"Taxe d'habitation"	www.impots.gouv.fr/portail/statistiques
Unemployment rate	At the "zone d'emploi" level 360 counties are divided into 7 "zone d'emplois".	www.insee.fr/fr/statistiques/1893230

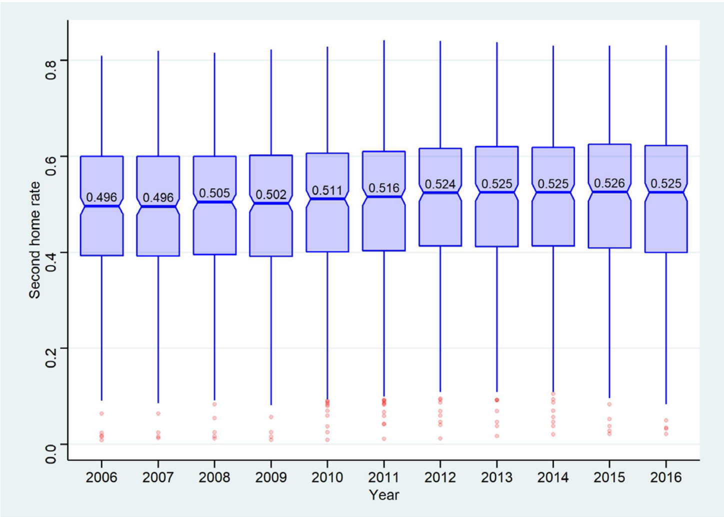

Figure A.1. Boxplot of the temporal trend in the raw rate of second homes as a proportion of the total number of houses. The marked number indicates the median of the annual second home rate among Corsican counties.

Table A.4. Estimates of the LCAR component (selected counties shown)

County ID	INSEE ID	Name	Mean estimates	SD
83	2A004	Ajaccio	-0.663	0.367
154	2A006	Alata	-1.088	0.109
52	2A017	Appietto	-0.260	0.107
267	2A032	Bastelicaccia	-0.239	0.152
80	2A041	Bonifacio	0.487	0.395
28	2A065	Cargese	0.892	0.139
30	2A090	Coggia	0.391	0.087
247	2A092	Conca	0.754	0.107

288	2A099	Cozzano	-0.109	0.104
257	2A131	Guagno	-0.473	0.093
248	2A139	Lecci	1.339	0.103
206	2A154	Marignana	0.795	0.086
209	2A196	Orto	-0.201	0.097
261	2A271	Sarrola-Carcopino	-0.888	0.111
16	2A300	San-Gavino-di-Carbini	0.349	0.135
207	2A348	Vico	0.970	0.089
250	2A362	Zonza	0.916	0.165
148	2B007	Albertacce	0.715	0.095
175	2B023	Asco	0.751	0.084
360	2B033	Bastia	-2.074	0.365
218	2B037	Biguglia	-2.813	0.136
316	2B047	Calacuccia	0.803	0.134
73	2B050	Calvi	0.910	0.203
54	2B073	Casamaccioli	0.779	0.113
300	2B095	Corscia	0.642	0.124
195	2B096	Corte	-0.742	0.253
299	2B120	Furiani	-2.674	0.124
298	2B147	Lozzi	0.496	0.097
68	2B150	Lumio	1.717	0.146
314	2B167	Montegrosso	0.752	0.102
97	2B289	Soveria	-0.018	0.189
119	2B305	San-Martino-di-Lota	-0.849	0.123
117	2B309	Santa-Maria-di-Lota	-1.033	0.136
27	2B329	Tralonca	-0.049	0.187
355	2B353	Ville-di-Pietrabugno	-1.018	0.145

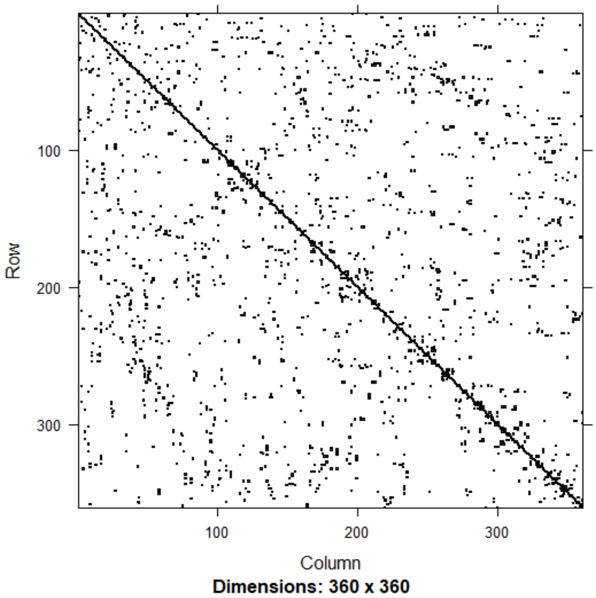

Figure A.2. Adjacency matrix: rows and columns identify areas; squares identify neighbors.

Appendix A.2: Technical details concerning econometric implementation

Appendix A.2.1 A typical representation of Bayesian hierarchical models

The Bayesian hierarchical model is usually specified as follows:

- Level 1 – Data model: [data | processes, parameters]
- Level 2 – Process model: [process | parameters]
- Level 3 – Parameter model: [parameters]

The first level describes the distribution of observed data, also known as the likelihood. The second level specifies the true latent process, given process parameters. For example, a spatial latent process is delineated by spatial random effects with precision parameters. At the third level, we assign hyperpriors to all parameters at the other levels. Finally, together with the likelihood, process and prior distribution of all parameters, we estimate the posterior distribution of the model parameters via Bayes theorem.

Appendix A.2.2 Model assessment criteria

The Deviance Information Criterion (DIC), the mean logarithmic conditional predictive ordinate (LCPO) and the root mean square error (RMSE) are employed in our study.

The DIC (Spiegelhalter, Best, Carlin, & Van Der Linde, 2002) is based on the posterior distribution of the deviance statistic and is defined as $\mathrm{DIC} = \overline{D} + p_D$, where $\overline{D}$ is the posterior expectation of the deviance and p_D is the effective number of parameters.

Alternatively, the CPO (Pettit, 1990) is used as a predictive measure of a Bayesian model. The CPO pertains to leave-one-out cross-validation and is defined as $\mathrm{CPO}_{it} = \pi\left(y_{it,obs} | y_{-(it),obs}\right)$, where $\pi\left(y_{it,obs} | y_{-(it),obs}\right)$ is the cross-validated predictive density at the omitted observation it given all the other data. Following the suggestion from (Roos & Held, 2011), we calculate the mean logarithmic CPO score, $\mathrm{LCPO} = -\frac{1}{N \times T} \sum_{t=1}^{T} \sum_{i=1}^{N} \log\left(\mathrm{CPO}_{it}\right)$.

Lastly, a simulation study is conducted by holding out the data for the most recent year. We then computed predicted values for the holdout units through the models trained by a training dataset. The root mean square error (RMSE) is considered to measure the closeness between the predicted second home incidence rate $\widehat{\pi_{it}}$ and the observed rate π_{it} and defined by $\mathrm{RMSE} = \sqrt{\frac{1}{N \times T} \sum_{t=1}^{T} \sum_{i=1}^{N} \left(y(s_i, t) - \widehat{y(s_i, t)}\right)^2}$.

Appendix A.3: Robustness Check

Concern may arise from three major issues: sensitivity to the priors, necessity of including covariates and endogeneity.

Regarding the prior sensitivity, we test different prior distributions to assess the change in the posterior distribution of all covariates and variance parameters. The tested priors are shown in Table A.5. For the fixed effects shown in Table A.6, the posterior distribution of all covariates obtained from the tested priors is almost the same as the posterior distribution of covariates using the default prior. In addition, the posterior distribution of the variance parameters is quite similar for the different priors. These results suggest that Model 5 should not be sensitive to priors.

To evaluate the need to include all covariates, we rerun Model 5 without any covariates, named as the convolution model. From Model 5 to the convolution model, the decrease of the DIC score can be clearly seen in Table A.7. We also notice that the posterior mean of the mixing parameter λ equals 0.544 ($95\% CI$, 0.348; 0.742) in the convolution model. The increase of the posterior mean of the mixing parameter demonstrates that spatially-referenced covariates capture some spatial variability. Rao (2003) stated that incorporating covariates to small area estimation models could increase the model predictive power. Our finding provides evidence for this point of view.

Lastly, the household variable may experience an endogeneity issue due to reverse causality. Since it lacks appropriate instruments in the context and the instrumental variable method within the Bayesian framework is still under investigated, we test strict exogeneity of the household variable via the Wooldridge's approach (Wooldridge, 2010). The lead household variable is included in the model additionally. We initially run Model 5 including the logged households as a linear predictor. Then, the lead-1 or lead-3 households is included in Model 5 additionally. The posterior estimates for the two additional variables are 0.016 $(95\%CI, -0.091; 0.124)$ and 0.051 $(95\%CI, -0.016; 0.118)$. Such result shows that there is not any endogenity issues (See Table A.8 for more details).

Table A.5. Tested hyperpriors in the prior sensitivity analysis

Component	Default	Test 1	Test 2
Spatially joint component (Γ)	$\log\tau \sim \text{logGamma}(1{,}5 \times 10^{-5})$	$\log\tau \sim \text{logGamma}(1{,}1 \times 10^{-4})$	$\sigma \sim Uniform(0, \infty)$
Temporally Structured component (Z)	$\log\tau \sim \text{logGamma}(1{,}5 \times 10^{-5})$	$\log\tau \sim \text{logGamma}(1{,}1 \times 10^{-4})$	$\sigma \sim Uniform(0, \infty)$
Space-time interaction term (Δ)	$\log\tau \sim \text{logGamma}(1{,}5 \times 10^{-5})$	$\log\tau \sim \text{logGamma}(1{,}1 \times 10^{-4})$	$\sigma \sim Uniform(0, \infty)$

$*\sigma = \frac{1}{\sqrt{\tau}}$

Table A.6. Estimated posterior mean, standard deviation and quantiles of the parameters for different hyperpriors

Prior	$\log\tau \sim \text{logGamma}(1{,}1 \times 10^{-4})$				$\sigma \sim Uniform(0, \infty)$			
	mean	SD	0.025quant	0.975quant	mean	SD	0.025quant	0.975quant
Intercept	-3.153	0.500	-4.156	-2.188	-3.223	0.522	-4.272	-2.217
physical landscapes	0.082	0.027	0.030	0.135	0.081	0.027	0.028	0.133
cultural landscapes	-0.047	0.019	-0.085	-0.010	-0.046	0.019	-0.084	-0.009
coastal county	0.720	0.109	0.506	0.935	0.722	0.110	0.505	0.939
mountainous county	0.217	0.083	0.054	0.379	0.218	0.083	0.054	0.381
$\sigma^2_{log_2(household)}$	0.217	0.039	0.151	0.304	0.227	0.040	0.157	0.315
log_2(interest rate)	-0.161	0.027	-0.217	-0.109	-0.165	0.033	-0.232	-0.102
log_2(council tax)	-0.015	0.011	-0.037	0.006	-0.018	0.011	-0.039	0.003
log_2(unemployment rate)	-0.049	0.044	-0.136	0.036	-0.056	0.044	-0.144	0.031
log_2(dis_gates)	0.051	0.057	-0.059	0.164	0.051	0.058	-0.062	0.168

log_2(dis_university)	0.170	0.069	0.038	0.308	0.173	0.071	0.036	0.316
σ_Γ^2	0.7355	0.1291	0.5207	1.0263	0.7756	0.1371	0.5403	1.0771
λ	0.2980	0.0903	0.1501	0.4998	0.3205	0.0937	0.1592	0.5217
Z	0.0015	0.0009	0.0004	0.0039	0.0026	0.0018	0.0006	0.0072
Δ	0.0093	0.0006	0.0082	0.0105	0.0093	0.0006	0.0082	0.0105

Table A.7. Refitted models assessment

	DIC	LCPO	RMSE
Model 5	27857.78	3.629	0.0341
Convolution Model 5	28226.61	3.711	0.0351

Table A.8. Estimated posterior mean and quantiles of the covariates for the strict exogenity test.

	mean	SD	0.025quant	0.975quant	mean	SD	0.025quant	0.975quant	mean	SD	0.025quant	0.975quant
Intercept	0.088	0.483	-0.870	1.030	-0.181	0.518	-1.211	0.828	-0.025	0.624	-1.274	1.181
physical landscapes	0.091	0.024	0.043	0.138	0.090	0.024	0.043	0.137	0.088	0.024	0.040	0.135
cultural landscapes	-0.028	0.017	-0.061	0.005	-0.031	0.017	-0.064	0.002	-0.039	0.017	-0.072	-0.006
coastal county	0.733	0.099	0.538	0.929	0.719	0.100	0.523	0.916	0.675	0.102	0.474	0.876
mountainous county	0.243	0.076	0.093	0.392	0.245	0.076	0.095	0.395	0.257	0.077	0.106	0.408
$log_2(household_t)$	-0.342	0.018	-0.378	-0.307	-0.350	0.055	-0.458	-0.241	-0.358	0.034	-0.425	-0.291
$log_2(household_{t+1})$					0.016	0.055	-0.091	0.124				
$log_2(household_{t+3})$									0.051	0.034	-0.016	0.118
log_2(interest rate)	-0.179	0.028	-0.237	-0.125	-0.211	0.037	-0.288	-0.140	-0.103	0.064	-0.231	0.022
log_2(council tax)	-0.009	0.011	-0.031	0.013	-0.004	0.011	-0.027	0.018	0.0003	0.012	-0.024	0.024

log_2(unemployment rate	-0.066	0.046	-0.157	0.023	-0.085	0.052	-0.188	0.018	-0.128	0.067	-0.261	0.002
log_2(dis_gates)	-0.001	0.051	-0.103	0.099	0.005	0.074	-0.141	0.151	0.029	0.077	-0.121	0.180
log_2(dis_university)	0.130	0.062	0.007	0.255	0.186	0.091	0.008	0.367	0.184	0.094	0.0004	0.371

References

Cressie, N. (1993). *Statistics for spatial data* (revised). New York: Wiley.

Pettit, L. (1990). The conditional predictive ordinate for the normal distribution. *Journal of the Royal Statistical Society: Series B (Methodological)*, *52*(1), 175–184.

Rao, J. N. K. (2003). *Small Area Estimation*. New York: Wiley.

Roos, M., & Held, L. (2011). Sensitivity analysis in Bayesian generalized linear mixed models for binary data. *Bayesian Analysis*, *6*(2), 259–278.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. Boston: The MIT Press.