The New England Journal of Statistics in Data Science Volume 0, 1–9 (2026)
DOI: https://doi.org/10.51387/26-NEJSDS112

Modeling First-Year Engineering Student Performance During the Pandemic

Ronny VALLEJOS*, Clemente FERRER, Andrea VÁSQUEZ, Rodrigo DÍAZ, and Teresita MARÍN

Abstract

Prompted by social upheaval in 2019, Chile initiated its first foray into entirely online education. Subsequently, owing to the global pandemic in 2020, all activities underwent an unavoidable shift to digital platforms, both on a global scale and within the university context. In response to this novel distance-learning system, the university encountered and surmounted new challenges during the four semesters that were conducted in this modality. This paper examines from a spatial statistics perspective the academic performance of first-year students enrolled in basic science courses across all engineering majors at Universidad Técnica Federico Santa María (USM) in Chile. This article emphasizes students' performance as a georeferenced variable in space and compares it with a regular semester in which lectures are conducted in a face-to-face format. In particular, we discuss (a) the spatial patterns observed in the two largest cities, (b) the social variables that are pertinent for the study, and (c) the quality and performance of conditionally autoregressive (CAR)-type processes in modeling multivariate lattice data. We also reflect on the opportunity to enhance the learning experience linked to the retention rate of freshman engineering students.

Keywords and phrases: First-year courses, COVID-19 pandemic, Spatial autocorrelation, Multivariate CAR process (MCAR).

1. INTRODUCTION AND MOTIVATION

The global COVID-19 pandemic disrupted the daily routines of every member of the worldwide educational community. The transition to online learning was a new and unexpected experience for both students and teachers alike, posing unique and complex challenges for education [9]. In Chile, the pandemic also had a significant impact on education. When schools and universities closed, traditional teaching was disrupted, prompting educational institutions to seek new ways to impart knowledge and assess their students.

At that time, several concerns arose regarding the education process. For instance, one of the most significant concerns that emerged was how the new remote learning modality would affect student performance. With disruptions in learning, there was a fear that grades would not accurately reflect learning, and students might struggle to keep up with their academic programs. On the other hand, the lack of personal interaction between teachers and students could impact the retention rate in higher education, especially in the early years, for engineering students, especially those who typically face challenges with foundational science courses [20]. In addition, the digital divide posed a challenge for both teachers and students, particularly for those without internet access or suitable devices in their homes. This impact was felt most significantly by low-income students and those residing in rural or remote areas.

The performance of first-year college students has been meticulously addressed by universities, as it is crucial for achieving a high retention rate and subsequent success in later courses within their academic programs [3, 15, 13]. USM accepts approximately 4,600 new students every year across all campuses in Chile, making it the largest engineering school in the country. Over the past decade, the university has implemented a series of methodological changes in the courses offered to freshman students. One of the key aspects that has been successful is the shift from traditional to active learning courses offered by the physics, chemistry, and mathematics departments. In some cases, both types of courses are offered, allowing students to choose in which ones they will enroll. All these good practices were drastically stopped in October 2019 because of the Chilean social strike and later again in March 2020 because of the COVID-19 pandemic. Despite the inconveniences experienced at that time, the constant practice of changing teaching techniques prior to the lockdowns led to the quick adaptation of remote teaching methods.

During the pandemic, most studies have focused on helping institutions adapt to emergency remote teaching. Studies to enhance teaching skills or student adaptation emerged after the pandemic had passed. Our study is meant to fill this gap.

*Corresponding author.

The database containing information about freshman students at our university includes georeferenced variables indicating the counties where they lived during the pandemic. This information is integrated with the social variables provided by the government for each county in Chile. The analytical framework employed in this work is known in spatial statistics as areal data, where the values of a georeferenced variable are associated with the entire area rather than a specific point. This approach forms the basis for our analysis, incorporating several areal variables in a multivariate setup for specific regions or counties in Chile. Spatial analysis can assist in addressing fundamental questions, such as whether nearby counties are more closely associated than those farther apart. It can also address more complex inquiries, such as predicting the performance of students in an unobserved specific area. Given our multivariate approach, addressing these questions requires considering the autocorrelation of each marginal process and the cross-correlation between each pair of processes. Furthermore, there is interest in selecting variables that exert a significant effect on the response.

This paper introduces a quantitative study on the academic performance of freshman students, considering the final scores obtained during the pandemic and social variables associated with each county where the students were located at that time. In Section 2, we provide a brief overview of the spatial statistics models that are essential to this study. Section 3 focuses on a particular study conducted using a database collected at USM that includes the recorded final grades of freshman students and relevant social variables. This section further provides a detailed description of the adjusted model and its estimates. Prediction maps are generated using three distinct spatial models, all estimated through a Bayesian perspective. The paper concludes with a discussion of the primary findings and practical considerations for future enhancements in the educational landscape of the university system. All the technical details, additional figures, and tables have been added to the Supplementary Material accompanying this paper.

2. PRELIMINARIES AND BACKGROUND

2.1 Analysis of Lattice Data

One of the fundamental properties of spatial data is the property of spatial dependence, which states that values that are close together in geographical space tend to be more similar than data values that are further apart. This property is known as Tobler's law [23]. It can be expressed in a slightly different way when the observations are associated with an area (lattice data), such that an observation of a variable at location i carries some information about what is observed for the same variable in areas that are close to i [11]. The way to include the information of the neighbors is through a contiguity matrix, which states the neighbor structure of the sites where the variable of interest is observed.

We consider the final grades of the freshman students in each course as lattice data. To propose a multivariate model for the scores, some basic areal models are reviewed in the next section.

2.2 CAR and MCAR Processes

Let $S \subset \mathbb{R}^2$, and consider a finite partition of S given by the set $\{S_1, \ldots, S_n\}$, i.e., $S = \bigcup_{i=1}^n S_i$ and $S_i \cap S_j = \emptyset$ for all $i \neq j$. The elements of the partition are referred to as areal units and are linked to a set of responses $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$.

The spatial variation in the response is modeled using covariates and a spatially structured random effect component $\boldsymbol{\Phi} = (\Phi_1, \ldots, \Phi_n)^\top$, where each Φ_i represents the spatial random effect for areal unit S_i. A widely used approach to model these random effects is to define the value at S_i conditionally on a linear combination of values from neighboring units.

Introduced by [4], CAR processes are commonly employed to capture this spatial dependence. A CAR model is specified via its full conditional distributions as:

$$\Phi_i \mid \boldsymbol{\Phi}_{-(i)} \sim \mathcal{N}\left(\rho \sum_{j \in \partial_i} b_{ij} \Phi_j, \tau_i^{-1}\right), \quad i = 1, \ldots, n, \tag{2.1}$$

where $j \in \partial_i$ denotes the indices of areal units that share a common boundary with S_j, the b_{ij} determine the strength of spatial interaction, ρ is a spatial autocorrelation parameter, τ_i is the conditional precision.

By the Hammersley-Clifford theorem and Brook's lemma, the set of full conditional distributions in (2.1) uniquely defines the joint distribution,

$$\boldsymbol{\Phi} \sim \mathcal{N}_n(\boldsymbol{0}, \{\boldsymbol{D}_\tau(\boldsymbol{I}_n - \rho \boldsymbol{B})\}^{-1}),$$

where $\boldsymbol{B} = \{b_{ij}\}_{i,j=1}^n \in \mathbb{R}^{n \times n}$ satisfies $b_{ii} = 0$, and $\boldsymbol{D}_\tau = \mathrm{diag}\{\tau_1, \ldots, \tau_n\}$.

Typically, the neighborhood structure is encoded through a binary contiguity matrix $\boldsymbol{W} = \{w_{ij}\}_{i,j=1}^n \in \mathbb{R}^{n \times n}$, where $w_{ij} = 1$ if areal units S_i and S_j share a common boundary, and $w_{ij} = 0$ otherwise. It is also conventional to set $w_{ii} = 0$. The corresponding diagonal matrix of neighborhood counts is defined as $\boldsymbol{D}_w = \mathrm{diag}\{w_{1+}, \ldots, w_{n+}\} \in \mathbb{R}^{n \times n}$ where $w_{i+} = \sum_{j \neq i} w_{ij}$ represents the number of neighbors of S_i.

These definitions are commonly used to specify the matrices $\boldsymbol{B}$ and $\boldsymbol{D}_\tau$ as follows: $\boldsymbol{D}_\tau = \tau \boldsymbol{D}_w$ and $\boldsymbol{B} = \boldsymbol{D}_w^{-1} \boldsymbol{W}_1$, thereby providing the full specification of the CAR process.

[2] extended the univariate CAR model to the multivariate setting by considering $\boldsymbol{\Phi} = (\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_p) \in \mathbb{R}^{n \times p}$, where each $\boldsymbol{\Phi}_i = (\Phi_{1,i}, \ldots, \Phi_{n,i})^\top \in \mathbb{R}^n$ represents the spatial random effects for the i-th variable across all regions. Under suitable assumptions, the joint distribution of $\boldsymbol{\Phi}$ can be expressed as

$$\mathrm{vec}(\boldsymbol{\Phi}) \sim \mathcal{N}_{np}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\mathrm{MCAR}}),$$

where $\mathbf{\Sigma}_{\text{MCAR}} = \mathbf{\Lambda}^{-1} \otimes (\boldsymbol{D}_w - \rho \boldsymbol{W})^{-1}$. Following [12], $\mathbf{\Lambda} \in \mathbb{R}^{p \times p}$ can be interpreted as a precision matrix that encodes the conditional association between variables within the same areal unit, while the matrix $(\boldsymbol{D}_w - \rho \boldsymbol{W})^{-1}$ captures the spatial association between different areal units for a given variable. The exact form of $\mathbf{\Lambda}^{-1}$ is given in Web Appendix B.

This model is referred to as MCAR(ρ, $\mathbf{\Lambda}$). If ρ is fixed at 1, the resulting model becomes an intrinsic multivariate CAR model, denoted IMCAR($\mathbf{\Lambda}$).

[26] laid the initial groundwork for extending the above model to a multivariate framework based on an adaptation of the linear model of coregionalization (LMC) for areal data. Subsequently, [5] introduced a unifying framework known as the $\mathcal{M}$-model, which allows for distinct spatial autocorrelation parameters across the multivariate responses.

In the $\mathcal{M}$-model, the multivariate spatial effect is represented as $\mathbf{\Phi} = \mathbf{\Theta}\mathcal{M}$, where $\mathbf{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p) \in \mathbb{R}^{n \times p}$ is a matrix whose columns are mutually independent spatial processes $\boldsymbol{\theta}_i \sim \mathcal{N}_n(\mathbf{0}, (\boldsymbol{D}_w - \rho_i \boldsymbol{W})^{-1})$, each following a CAR model with its own spatial autocorrelation parameter ρ_i. The matrix $\mathcal{M} = \{m_{ij}\} \in \mathbb{R}^{p \times p}$ defines the loadings of the different underlying CAR spatial effects for each variable. Therefore, each spatial random vector effect can be written as:

$$\mathbf{\Phi}_j = \sum_{i=1}^{p} \boldsymbol{\theta}_i m_{ij}, \quad j = 1, \ldots, p.$$

2.3 Estimation and Mapping

The implementation of the IMCAR($\mathbf{\Lambda}$), MCAR($\rho, \mathbf{\Lambda}$), and $\mathcal{M}$-models is commonly conducted under a Bayesian framework due to the explicit conditional structure [10]. The computational implementation can be constructed, e.g., using `WinBUGS` [2] or the R package `INLAMSM` [16], which allows the implementation of the models discussed in Section 2.2.

For lattice data, units in close proximity to each other with similar values exhibit a spatial pattern indicative of positive spatial autocorrelation. Identifying groups of units in close proximity with high values is often of particular interest because they suggest a cluster of elevated risk, possibly originating from a common source [19, p. 11].

In this work, we have chosen to use multivariate CAR models due to their compelling advantages for modeling multivariate areal data. These models jointly capture spatial dependence both within and across variables, and their sparse precision matrices allow for computational efficiency, making them particularly suitable for large datasets. Compared to alternatives such as multivariate Gaussian processes, MCAR models are more scalable while maintaining modeling flexibility. Moreover, they integrate seamlessly into hierarchical Bayesian frameworks, facilitating robust parameter estimation and uncertainty quantification in complex spatial settings.

3. SPATIAL MODELING OF ACADEMIC PERFORMANCE

3.1 The Dataset

The dataset employed in this study encompasses a compilation of academic records for all students enrolled at Universidad Técnica Federico Santa María from 2018 to 2022. We restricted our analysis to three years, as 2018 marked the implementation of a revised teaching methodology in first-year mathematics courses, rendering data from earlier years not directly comparable.

Between 2018 and 2022, several disruptions prevented these years from being treated as a continuous sequence. In 2018, a student strike halted teaching for nearly three months. In 2019, Chile experienced a social uprising that forced universities to switch to online instruction from October onward, with the fall semester extending into February 2020. During 2020–2021, lectures were delivered online due to the pandemic, and 2022 marked the return to regular in-person operations.

With multiple entries, each record in the dataset represents an individual instance of a student's academic performance in a specific subject, along with their personal and academic details.

The dataset was preprocessed to streamline its structure for a focused spatial and academic performance analysis. This included data cleaning and the exclusion of incomplete cases. A logarithmic transformation was applied to the USC variables to stabilize their variance, as their values ranged between 600 and 800. Additionally, since the social variables were scaled within the $(0, 1)$ interval, the transformation helped integrate all variables into a common analytical framework.

To examine differences in outcomes before, during, and after the pandemic, the first semesters of the years 2018, 2020, and 2022 were selected across the Valparaíso Region (**V**) and the Santiago Metropolitan Region (**RM**). The choice of these regions was driven by their having the largest student populations as can be seen in Table 1, and hosting the principal campuses of the university. In addition, other regions exhibit sparse enrollment data, often with several counties registering zero observations. It is important to highlight that the Valparaíso Region comprises 36 counties, whereas the Santiago Metropolitan Region consists of 52 counties.

In the Valparaíso Region, there were 689 students in 2018, 616 in 2020, and 567 in 2022. Similarly, in the Santiago Metropolitan Region, there were 881 students in 2018, 1010 in 2020, and 898 in 2022. A detailed breakdown of first-year student counts in these two regions is available in Tables 6 and 7 of the Supplementary Material.

Temporal effects were not incorporated into the analysis for two main reasons. First, the available data spans only a limited number of academic years (2018, 2020, and 2022),

Table 1. Frequency of first-year students by year and region.

	Region															
Year	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIV	XV	XVI	RM
2018	44	47	37	125	689	162	66	13	37	66	9	20	11	37	12	881
2020	38	49	44	107	616	123	73	15	29	52	8	10	16	29	11	1010
2022	36	38	24	84	567	114	41	5	23	46	7	9	15	19	11	898

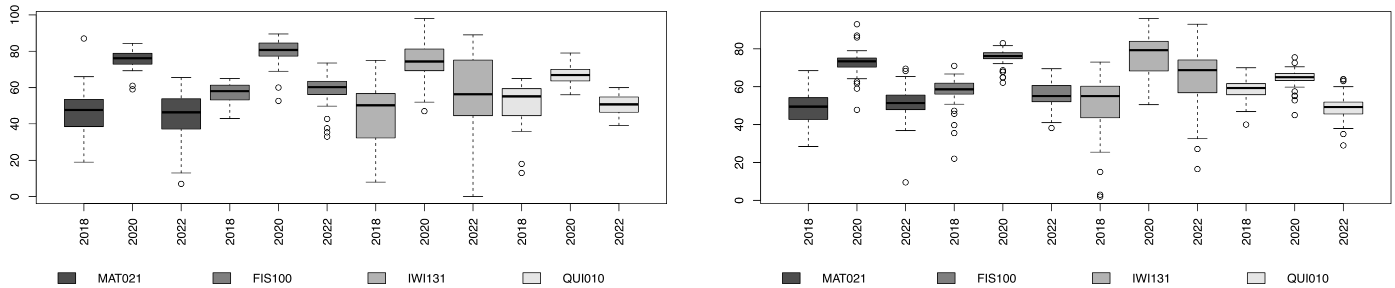

Figure 1: Student grades for the years 2018, 2020, and 2022 for the MAT021, FIS100, IWI131, and QUI010 subjects in the Valpariso Region (Left) and the Santiago Metropolitan Region (Right).

which is insufficient to reliably capture long-term temporal patterns. Second, the student population changes from year to year, with each cohort bringing different characteristics and backgrounds. This inter-cohort variability introduces additional complexity that makes it difficult to isolate temporal effects in a consistent and interpretable way. Therefore, we chose to focus the analysis on spatial patterns within each year, where the assumptions and comparisons are more stable and meaningful.

The grades (recorded on a 0–100 scale) obtained by first-year students in the Mathematics I (MAT021), Introduction to Physics (FIS100), Programming (IWI131), and Chemistry and Society (QUI010) courses were averaged by county across the Valparaíso Region and Santiago Metropolitan Region (refer to Tables 2 and 3 in the Supplementary Material).

Each record now directly links academic performance in specific subjects with spatial data. Additionally, in pursuit of fitting an appropriate model, various economic, social, and educational variables that might explain the grades obtained in each county are considered in an auxiliary dataset (refer to Tables 4 and 5 in the Supplementary Material). These are as follows:

1. **Poverty Index (PI)**: The Social Development and Family Ministry compiles and publishes estimates of the poverty rate at the county level, considering advanced methodologies that combine survey information and administrative records.
2. **Municipal Development Index (MDI)**: Provided by the Chilean Institute of Municipal Studies (ICHEM) and the Institute for Habitat Studies (IEH), both of which are parts of Universidad Autónoma, the MDI is a composite index constructed to represent information from various sources, aiming to measure different aspects of reality at the municipal level. This reflects high rates of inequality in municipal development across Chile.
3. **Average University Selection Test (USC) Score**: The USC is a standardized written test implemented in Chile for the university admission process. The USC score is the average calculated from the scores obtained in the language and communication and mathematics tests.
4. **Rate of Private Schools (RPS)**: Based on reports from the Library of the National Congress of Chile, the rate of private schools is calculated as the ratio between the number of private schools and the total number of educational establishments in the counties.

3.2 Exploratory Data Analysis

The observed patterns (see Figure 1) in grades are similar in Santiago and Valparaíso, with the first term of 2020 yielding the highest scores in all subjects. This highlights that during the first semester of the pandemic, the average scores were significantly greater than those in the first semester before and after the pandemic.

This hypothesis has been confirmed through multiple tests (details not reported here). We also observe that on both campuses, the best performance is in Introduction to Physics, while Programming and Mathematics I show the lowest performance. The passing threshold for a course is set at 55 out of 100. In terms of score stability over the years, Chemistry and Society is the strongest course.

In Figure 2, the maps displaying the final grades in each course for the Valparaíso and Santiago regions are presented. In cases where data are unavailable for a specific county, the respective areal unit is colored orange, and the missing value is imputed using the average of the other counties

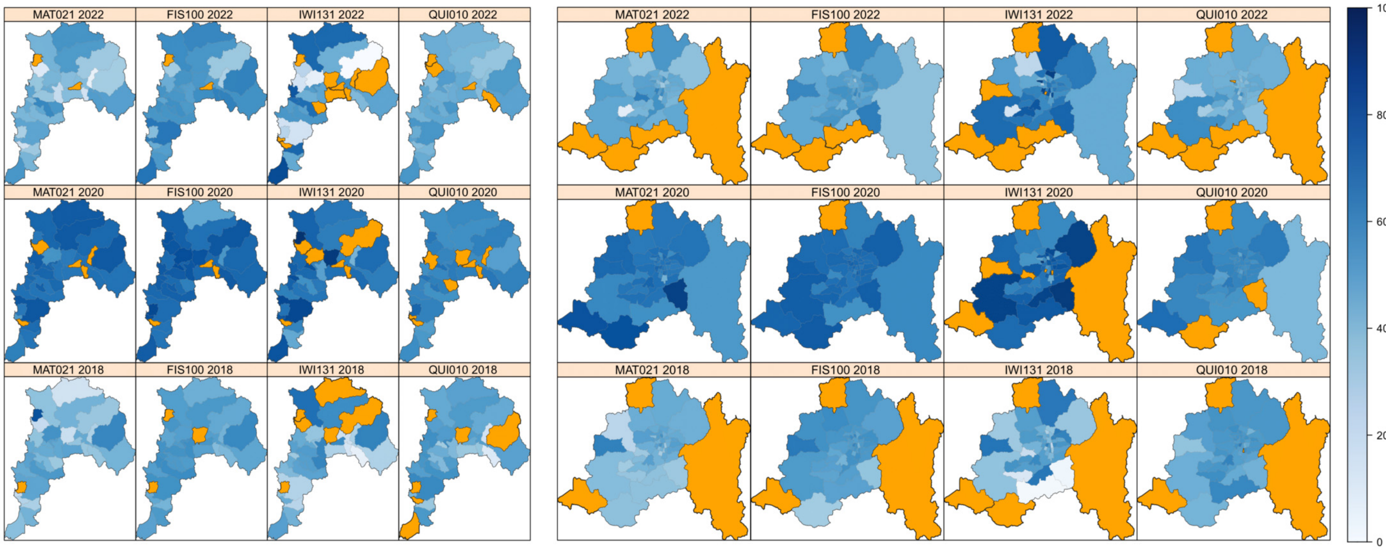

Figure 2: Spatial analysis of student grades for the years 2018, 2020, and 2022 for the MAT021, FIS100, IWI131, and QUI010 subjects in the Valparaíso Region (Left) and the Santiago Metropolitan Region (Right).

to avoid bias. Visually, it is evident that in 2020, the grades are higher than those in both the prepandemic and postpandemic periods. Various explanations have been proposed for this phenomenon, making it challenging to investigate from a scientific perspective. For instance, the limited experience of professors with online teaching methods may have led to the use of inadequate evaluation methods in this context. Additionally, the unintended collaboration among student groups during midterms and exams, which involves sharing crucial information, has been identified as a significant factor contributing to unprecedented approval rates. Despite the challenges in providing a rigorous explanation for this behavior, the data underscore an existing pattern that calls for further analysis.

According to Moran's index (Anselin, 1995 [1]), significant autocorrelation exists for most of the scoring variables examined in the study (refer to Web Appendix A). Consequently, it is worthwhile to develop spatial models that account for spatial interactions both between and within variables.

3.3 Model Specification

To incorporate the covariates described in Section 3.1, we model the average student performance per county and year using a Bayesian hierarchical spatial model. Let Y_{it} denote the average score for county $i = 1, \ldots, n$ at year $t = 1, \ldots, p$. In our study, the number of years is $p = 3$ and the number of counties is $n = 36$ for the Valparaiso Region and $n = 52$ for the Santiago Metropolitan Region.

We assume a Gaussian likelihood $\log(Y_{it}) \sim \mathcal{N}(\mu_{it}, \tau^{-1})$, with mean $\mu_{it} = \alpha_t + \mathbf{x}_{it}^{\top}\boldsymbol{\beta}_t + \phi_{it}$, where α_t is a year-specific intercept, $\mathbf{x}_{it} \in \mathbb{R}^q$ is the vector of $q = 4$ covariates (PI, MDI, USC, RPS), $\boldsymbol{\beta}_t \in \mathbb{R}^q$ is the vector of regression coefficients that vary by year, and ϕ_{it} denotes spatially structured random effects modeling residual spatial variation.

The spatial random effects ϕ_{it} for all counties and years are modeled jointly as $\mathrm{vec}(\boldsymbol{\Phi}) \sim \mathcal{N}_{np}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is specified according to the spatial models described in Section 2.2

Inference is performed using INLA [18]. Further details on priors and posterior inference are provided in the Supplementary Material (Web Appendix C). In particular, Figure 1 and Table 1 in the Supplementary Material present forest plots with 95% credible intervals for each course, method, year, and area. These plots provide an overall perspective on the dispersion associated with the variables considered.

As a sensitivity analysis, we refitted the MCAR model under a likelihood scaled by the county–year sample size n_{it}. The fixed-effect posterior means differ by at most 0.06 on the log scale, and credible-interval widths differ by less than 2%, indicating that the conclusions are robust to sample-size weighting (Web Appendix F of the Supplementary Material).

The full posterior summaries for all fitted models are reported in the Supplementary Material (Tables 8 and 9). According to the Deviance Information Criterion (DIC) [21] and the Widely Applicable Information Criterion (WAIC) [24, 25], where lower values indicate better fit (see also Tables 10 and 11), the MCAR model provides the best overall fit, supporting the estimation of the spatial autocorrelation parameter rather than fixing it. Consequently, Figure 3 shows the fitted maps for the four subjects and the three years using the MCAR specification only.

To assess model adequacy, we inspected standardized predictive residuals from the MCAR specification. Figure 4

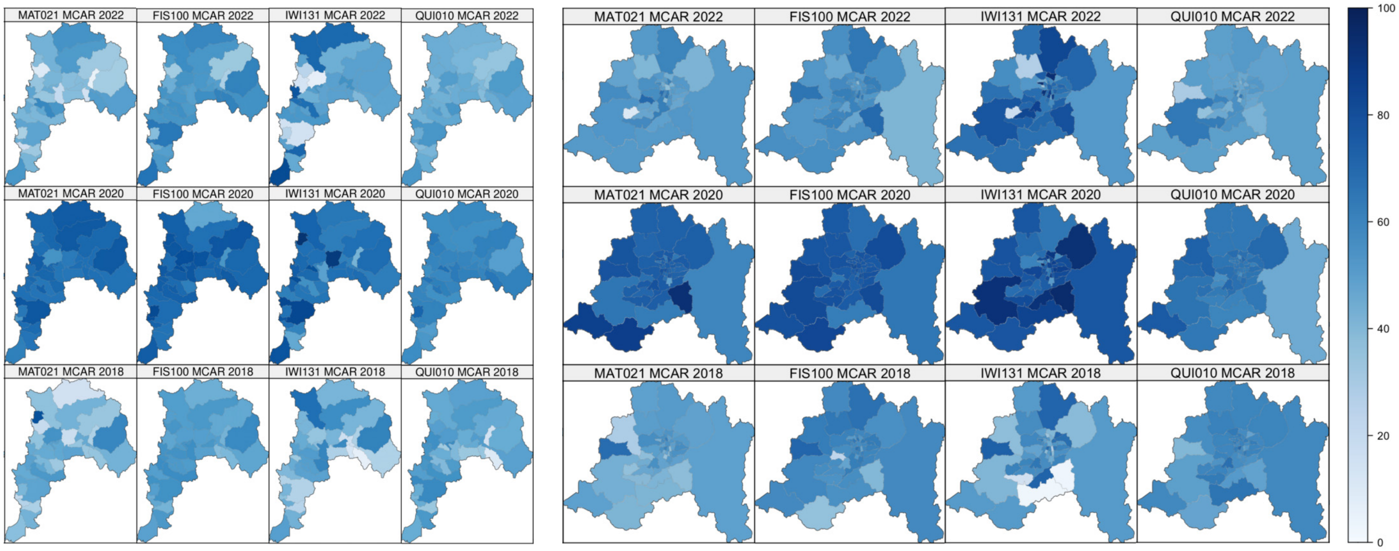

Figure 3: MCAR fitted maps of county-level average grades for 2018, 2020, and 2022 in the Valparaíso Region (left) and the Santiago Metropolitan Region (right), for MAT021, FIS100, IWI131, and QUI010.

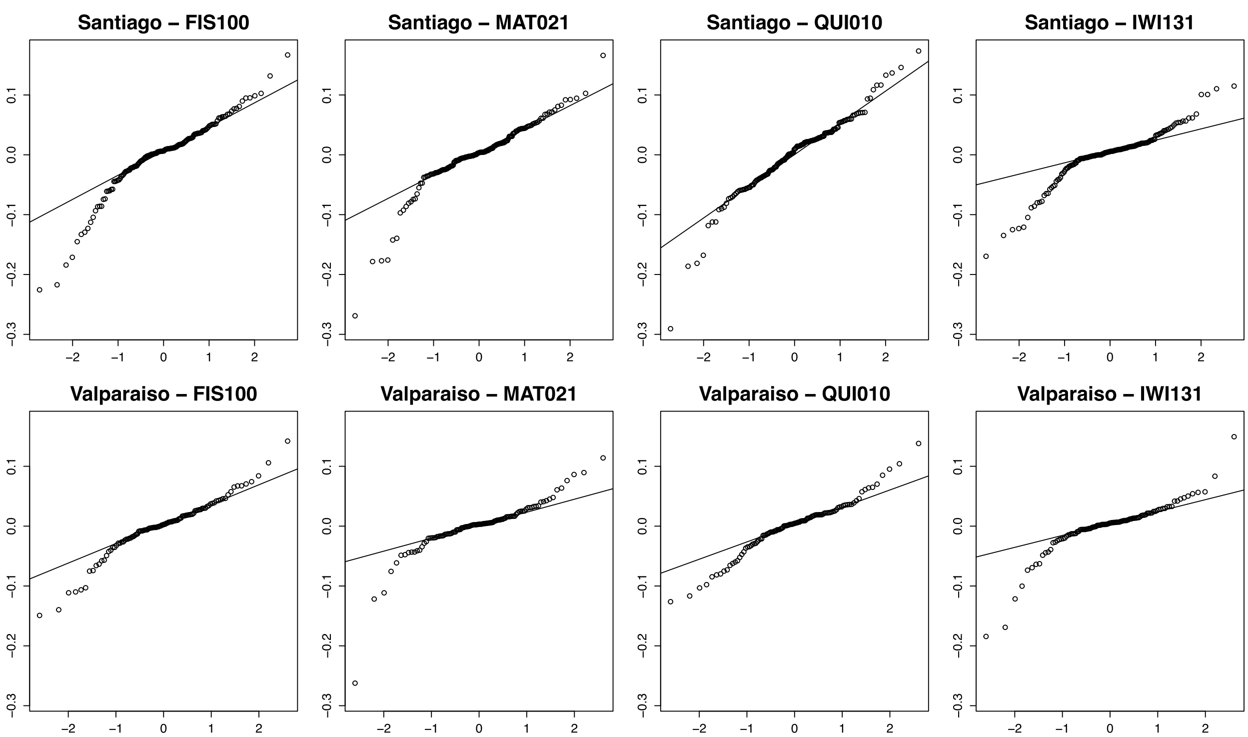

Figure 4: Normal QQ-plots of standardized predictive residuals from the MCAR model, computed on the log-response scale, for the four subjects in the Valparaíso Region and the Santiago Metropolitan Region.

presents QQ-plots based on residuals computed on the log-response scale for the four subjects in both regions. The plots indicate an overall reasonable agreement with normality, especially in the central part of the distribution, although some noticeable deviations are present in the tails.

Across the four covariates, only one association has a credible interval that excludes zero: in Santiago, the municipal development index is positively associated with grades in MAT021 in 2018, with $\hat{\beta} = 0.77$ and a 95% credible interval of $[0.15, 1.40]$. On the original grade scale, this corresponds to roughly an 8% higher county-average grade per 0.1 increase in the index. The remaining patterns are

Table 2. Computational times, measured in seconds, of each model by subject and study region.

	MAT021			FIS100		
	IMCAR	MCAR	$\mathcal{M}$	IMCAR	MCAR	$\mathcal{M}$
Valparaíso Region (s)	2.20	2.38	4.78	2.10	2.26	5.32
Santiago Metropolitan Region (s)	2.30	2.46	6.34	2.36	2.32	5.67

	IWI131			QUI010		
	IMCAR	MCAR	$\mathcal{M}$	IMCAR	MCAR	$\mathcal{M}$
Valparaíso Region (s)	2.16	2.39	7.08	2.08	2.35	8.27
Santiago Metropolitan Region (s)	2.20	2.45	6.65	2.30	2.39	6.18

weaker. The admission score is positively associated with grades in IWI131 in Santiago across the three years, with point estimates near 0.8 but credible intervals that always include zero. The poverty index changes sign across courses and years, with large but uncertain magnitudes; the rate of private schools shows no stable pattern. Most estimates therefore carry substantial posterior uncertainty, and the associations above should be read as suggestive rather than conclusive.

The change in the sign of the PI coefficient between the lockdown semester (2020) and the post-pandemic semester (2022), particularly for IWI131 in Valparaíso, is consistent with the temporary buffering role of pandemic-related social policies documented across Latin America [22, 6]. With the data at hand we cannot test this mechanism formally, and we present this only as one possible reading of the observed pattern.

These findings suggest the value of admissions policies that adopt a more holistic approach, taking into account the disparities that students endure because of their socioeconomic background. Universities could explore initiatives to implement useful indicators based on socioeconomic factors for freshman students to mitigate the disadvantages faced by these students, aligning with broader educational equity objectives. These new policies could complement existing initiatives driven by the government of Chile, such as the PACE[1] program. While the current focus of the latest initiatives has been on promoting gender equality in the admission process, they should encompass socioeconomic considerations to broaden their impact.

4. DISCUSSION AND FINAL COMMENTS

This study examined first-year engineering student performance at Universidad Técnica Federico Santa María from a spatial statistics perspective, contrasting a pre-pandemic semester (2018), the lockdown period (2020), and a post-pandemic semester (2022). By treating county-level average grades as areal data and integrating socioeconomic covariates, the analysis provides a structured view of how spatial dependence and contextual factors relate to academic outcomes in core first-year courses.

The computational burden of the proposed models is manageable and does not require specialized hardware. All results were obtained using a Macbook Air M3 computer equipped with an Apple M3 processor and 16.0 GB of RAM, and the corresponding runtimes are summarized in Table 2.

Substantively, the estimated associations between socioeconomic covariates and grades are heterogeneous across regions, periods, and subjects, and most carry substantial posterior uncertainty. Only one coefficient (MDI for MAT021 in Santiago in 2018) has a credible interval that excludes zero. The remaining estimates should be read as contextual associations at the county level rather than as causal effects at the individual level, and as a description of the strongest patterns observed in the data rather than as evidence of stable mechanisms.

Taken together, these results are consistent with the use of socioeconomic information as a complement to academic data when supporting first-year students, while acknowledging the contextual nature of the associations and the substantial uncertainty in the estimated effects. In this regard, USM's ongoing digital transformation provides an opportunity to leverage institutional data and digital platforms to better address the educational challenges highlighted by this study, helping tailor support to students' diverse socioeconomic backgrounds and fostering a more inclusive learning environment.

DISCLOSURE STATEMENTS

No potential conflicts of interest are reported by the authors.

DATA AVAILABILITY STATEMENT

The participants of this study did not give written consent for their data to be shared publicly, so due to the sensitive nature of the research supporting data is not available. However, the code can be accessed at the following GitHub link.

[1] https://acceso.mineduc.cl/admision-universidades-2023/portal-pace/

SUPPLEMENTARY MATERIAL

The online Supplementary Material provides comprehensive details regarding the Moran index, the structure of matrix Λ, and the Bayesian inference framework employed in the estimation process. Additionally, we include the complete dataset used in the application, along with tables detailing supplementary results. Finally, this material features a dedicated section on the sensitivity study of the MCAR models.

ACKNOWLEDGMENTS

The authors express their gratitude to Felipe Osorio for engaging and insightful discussions and for revising a preliminary version of this manuscript. The authors are grateful to the Subject Editor and the three anonymous reviewers, whose comments substantially improved the clarity and quality of this work.

FUNDING

This work was supported by UTFSM through OEA2023 grant no. 232, and by Fondecyt grant no. 1230012.

Accepted 2 September 2026

REFERENCES

[1] Anselin, L. (1995). Local Indicators of Spatial Association—LISA. *Geographical Analysis* **27** 93–115.

[2] Banerjee, S., Carlin, B. and Gelfand, A. (2003) *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC, New York. MR5123534

[3] Barry, C. A. and Jones, A. L. (1959). A study of the performance of certain freshman students. *The Journal of Educational Research* **52** 163–166.

[4] Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B* **36** 192–225. MR0373208

[5] Botella-Rocamora, P., Martinez-Beneito, M. A. and Banerjee, S. (2015). A unifying modeling framework for highly multivariate disease mapping. *Statistics in Medicine* **34** 1548–1559. MR3334675

[6] Busso, M., González, M. and Scartascini, C. (2021). Social protection and informality in Latin America during the COVID-19 pandemic. *CEPAL Review* (133) 27–47.

[7] Cressie, N. (1993) *Statistics for spatial data*. John Wiley and Sons, New York. MR1239641

[8] ECLAC (2022). *Education during the pandemic: An opportunity to transform education systems in Latin America and the Caribbean*. United Nations Economic Commission for Latin America and the Caribbean.

[9] Flores, A., Capiello, L. and Salinas, I. (2023). Challenges and successes of emergency online teaching in statistics courses. *Journal of Statistics and Data Science Education*. (in press). https://doi.org/10.1080/26939169.2023.2231036.

[10] Gelfand, A. and Vounatsou, P. (2003). Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics* **4** 11–25.

[11] Haining, R. and Li, G. (2020) *Modelling Spatial and Spatial-Temporal Data: A Bayesian Approach*. CRC Press, Boca Raton.

[12] Jin, X., Banerjee, S. and Carlin, B. P. (2007). Order-Free Co-Regionalized Areal Data Models with Application to Multiple-Disease Mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69**(5) 817–838. https://doi.org/10.1111/j.1467-9868.2007.00608.x. MR2368572

[13] Lin, S., Mastrokoukou, S., Longobardi, C., Bozzato, P., Gastaldi, F. G. M. and Berchiatti, M. (2022). Students' transition into higher education: The role of self-efficacy, regulation strategies, and academic achievements. *Higher Education Quarterly* **77** 121–137.

[14] Londoño-Vélez, J. and Querubín, P. (2022). The Impact of Emergency Cash Assistance in a Pandemic: Experimental Evidence from Colombia. *American Economic Review* **112**(5) 1443–1474. https://doi.org/10.1257/aer.20201463.

[15] McCammon, S., Golden, J. and Wuensch, L. (1998). Predicting course performance in freshman and sophomore physics: Women are more predictable than men. *Journal of Research in Science Teaching* **25** 501–510.

[16] Palmi-Perales, F., Gómez-Rubio, V. and Martinez-Beneito, M. A. (2021). Bayesian Multivariate Spatial Models for Lattice Data with INLA. *Journal of Statistical Software* **98** 1–29.

[17] Ribeiro, F. G., Neri, M. et al. (2023). The impact of conditional cash transfers on poverty, inequality, and employment during COVID-19: Evidence from Brazil. *World Development* **161** 106033. https://doi.org/10.1016/j.worlddev.2022.106033.

[18] Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71** 319–392. https://doi.org/10.1111/j.1467-9868.2008.00700.x. MR2649602

[19] Schabenberger, O. and Gotway, C. (2005) *Statistical Methods for Spatial Data Analysis*. Chapman & Hall/CRC, Boca Raton. MR2134116

[20] Sithole, A., Chiyaca, E. T., McCarthy, P., Mupinga, D. M., Bucklein, B. K. and Missouri, J. (2017). Student attraction, persistence and retention in STEM programs: successes and continuing challenges. *Higher Educational Studies* **7** 46–59.

[21] Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and 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. MR1979380

[22] Stampini, M., Ibarrarán, P., Medellín, N., Villa, J. M. and Gargiulo, C. (2021). Adaptive, but not by design: Cash transfers in Latin America and the Caribbean before, during and after the COVID-19 pandemic. *International Social Security Review* **74**(3) 55–75. https://doi.org/10.1111/issr.12250.

[23] Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography* **46** 299–308.

[24] Watanabe, S. (2010). Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *Journal of Machine Learning Research* **11** 3571–3594. MR2756194

[25] Watanabe, S. (2013). A Widely Applicable Bayesian Information Criterion. *Journal of Machine Learning Research* **14** 867–897. MR3049492

[26] Xiaoping, J., Banerjee, S. and Carlin, B. (2007). Order-free co-regionalized areal data models with application to multiple-disease mapping. *Journal of the Royal Statistical Society Series B* **69** 817–838. MR2368572

Ronny Vallejos. Avenida España 1680, Valparaíso, Departamento de Matemática, Dirección de Educación a Distancia, Universidad Técnica Federico Santa María, Chile.
E-mail address: ronny.vallejos@usm.cl

Clemente Ferrer. Avenida España 1680, Valparaíso, Departamento de Matemática, Universidad Técnica Federico Santa María, Chile.
E-mail address: clemente.ferrer@usm.cl

Andrea Vásquez. Avenida España 1680, Valparaíso, Departamento de Informática, Universidad Técnica Federico Santa María, Chile.
E-mail address: andrea.vasquez@usm.cl

Rodrigo Díaz. Avenida España 1680, Valparaíso, Dirección de Educación a Distancia, Universidad Técnica Federico Santa María, Chile.
E-mail address: rodrigo.diaz@usm.cl

Teresita Marín. Avenida España 1680, Valparaíso, Dirección de Enseñanza y Aprendizaje, Universidad Técnica Federico Santa María, Chile.
E-mail address: teresita.marin@usm.cl