MODELLISTICA E CALCOLO SCIENTIFICO

MOX

MODELING AND SCIENTIFIC COMPUTING

MOX-Report No. 92/2023

Commuting time as a determinant of higher education students' performance: the case of Politecnico di Milano

Burzacchi, A.; Rossi, L.; Agasisti, T.; Paganoni, A. M.; Vantini, S.

Commuting time as a determinant of higher education students' performance: the case of Politecnico di Milano

Arianna Burzacchi*[a] ⓘ, Lidia Rossi*[a,b] ⓘ, Tommaso Agasisti[b] ⓘ, Anna Maria Paganoni[a] ⓘ and Simone Vantini[a] ⓘ

[a] MOX Laboratory, Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133, Milano, Italy;
[b] Department of Management, Economics and Industrial Engineering, Politecnico di Milano, Via Lambruschini 4/B, 20156, Milano, Italy

ABSTRACT
Despite its crucial role in students' daily lives, commuting time remains an under-explored dimension in higher education research. To address this gap, this study focuses on challenges that students face in urban environments and investigates the impact of commuting time on the academic performance of first-year bachelor students of Politecnico di Milano, Italy. This research employs an innovative two-step methodology. In the initial phase, machine learning algorithms trained on GPS data from anonymous users are used to construct accessibility maps to the university and to obtain an estimate of students' commuting times. In the subsequent phase, authors utilize polynomial linear mixed-effects models and investigate the factors influencing students' academic performance, with a particular emphasis on commuting time. Notably, this investigation incorporates a causal framework, which enables the establishment of causal relationships between commuting time and academic outcomes. The findings underscore the significant impact of travel time on students' performance and may support policies and implications aiming at improving students' educational experience in metropolitan areas. The study's innovation lies both in its exploration of a relatively uncharted factor and the novel methodologies applied in both phases.

KEYWORDS
Academic Performance; Accessibility Maps; Data-driven Approach; Generalized Propensity Score; GPS Data.

1. Introduction

Access to higher education is widely acknowledged as a key determinant of social mobility, economic advancement, and students' personal development. Universities are pivotal institutions that serve as centers of knowledge creation, skill development, and intellectual growth, playing a significant role in shaping societies, producing a qualified workforce, and promoting innovation. Within metropolitan areas, the concentration of universities presents both opportunities and challenges in terms of accessibility. Metropolitan regions are characterized by high population densities, diverse socio-economic profiles, and complex transportation networks. The availability and proximity of universities within these areas shape educational opportunities and may create disparities that impact student outcomes. Understanding the university accessibility

---
* The authors contributed equally to the work.
A. Burzacchi. Email: arianna.burzacchi@polimi.it; L. Rossi. Email: lidia.rossi@polimi.it

patterns and their implications is hence essential for promoting equitable opportunities in students' educational experiences.

Measuring accessibility consists of evaluating the ease with which individuals can reach and utilize university resources. Commuting time, as an indicator of proximity, offers valuable insights into the accessibility of universities within metropolitan areas (Geurs and van Wee 2004; Hu et al. 2023). It reflects the travel duration required for students to reach their educational institutions, taking into account factors such as distance, transportation modes, traffic congestion, and infrastructure quality. More specifically, commuting time is defined as the amount of time spent on one-way travel from the home location to the study or working institution. It has a central role in most lives of students and workers, both due to its daily repetition within the personal routine of individuals, and for the large amount of time taken away from everyday life. Eurostat (2020) reports the estimated average commuting time of employed people in 2019 in the EU-27 to be equal to 25 minutes per day. Regarding students, the estimate increases: according to Gwosc et al. (2021), in the period between 2018 and 2021, the median time of a commute for a European higher education student living in a student university residence was around 15 minutes and reached 40 minutes for students living with their parents.

Among contextual factors that could affect students' school performance (Hanushek 1979), commuting time is a key factor, and its investigation is the major goal of this research. The purpose of the study is to shed light on the relationship between commuting time and academic performance of higher education students. The case study of the research regards students who are enrolled in engineering bachelor degrees at Politecnico di Milano, a university institution with two campuses located in the city of Milan, Italy.

The research is carried out in a two-step procedure, starting from the estimation of students' commuting time, and moving to the analysis of its impact on students' academic performance. Students' travel time is estimated by means of the non-parametric statistical method of Kernel Regression Estimation using Global Positioning System (GPS) data of anonymous individuals. Here lies one of the main contributions of the research, which combines an innovative data source, i.e. GPS smartphone data, and a data-driven methodology for travel time estimation, overcoming the common limitations found by other studies. In the second phase, the impact of commuting time on academic performance is assessed through the application of balance weighting methods and multilevel polynomial regression models. Students' academic performance is measured through the Grade Point Average (GPA), defined as the average of the marks obtained by the students in the evaluation tests. Other students' information, such as basic generalities (e.g., age, gender) and school-specific features (e.g., bachelor track), are included in the model as confounders to control their effect on academic performances together with commuting time. A multilevel model is used to take into account the variability within different programs and balancing methods are modified accordingly, leading to a methodological innovation in this application field. According to the authors' knowledge, this is the first study that designs the analysis within a causal framework using balancing methods to evaluate the causal relationship between commuting time and students' performances. A graphical representation of the analytical framework is reported in Figure 1.

The results of this study have the potential to significantly influence the educational experience of students, particularly for those who frequently commute long distances to their campus. Indeed the research highlights the relevance of commuting duration in students' daily encounters and offers insights that could inform policies for the

2

enhancement of students' learning experience. It is to be noted that the whole analysis is scalable and can be straightforwardly extended to other case studies of university institutions located in other metropolitan cities.

The paper is organized as follows. A review of the literature is reported in Section 2, with both a description of the contextual framework of the research and the analysis of other works with the same goal. Then, methodology and data are explained in Section 3, while in Section 4 the results are described. Finally, Section 5 is dedicated to conclusions.

The analysis is carried out by means of the software R (R Core Team 2022) exploiting some relevant functions from the R packages `sf, np, lme4, lmerTest, WeightIt`, and `marginalEffects` (Arel-Bundock 2023; Bates et al. 2015; Greifer 2023; Hayfield and Racine 2008; Kuznetsova, Brockhoff, and Christensen 2017; Pebesma 2018). The code is available under request to the corresponding authors.
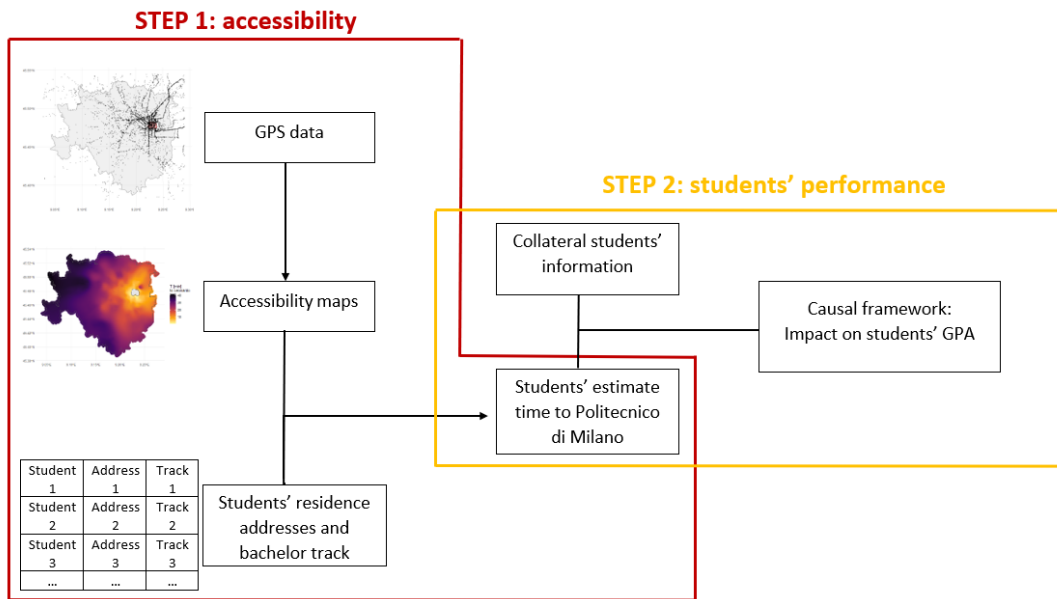

Figure 1. A graphical representation of the analytical framework

2. State of the art

2.1. *Conceptual framework - determinants of higher education students performance*

As theorized by Hanushek (1979), the factors that affect students' school performance are multiple. Based on multiple literature reviews (Abu Saa, Al-Emran, and Shaalan 2019; Hellas et al. 2018; Hu et al. 2017; Kumar, Singh, and Handa 2017; Lei and Li 2015; Rastrollo-Guerrero, Gómez-Pulido, and Durán-Domínguez 2020; Shahiri, Husain, and Rashid 2015), the determinants of university performance, as assessed by students' evaluations, can be grouped into three categories: personal, socio-economic, and academic factors.

The first set of factors that impact academic success pertains to the demographics of students, such as gender, age, and nationality. Such determinants are investigated in

observational studies from the literature like Alhajraf and Alasfour (2014); Danilowicz-Gösele et al. (2017); Yousef (2019). Personal factors also comprise Motivators, Attitudes, and Beliefs (MAB) factors like fear of punishment and distributive justice, as well as personality traits like the Big Five (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism), as explored in the works of Flom, Green, and Wallace (2021) and Corazzini et al. (2021) respectively.

Socio-economic parameters are the second set of factors that impact academic performance. This group pertains to factors related to the family, such as family size, family income, parents' education and occupation, and residence location. Works investigating the relationship between academic performance and such factors are, for instance, Silva, Vautero, and Usssene (2021); Rodriguez-Hernandez, Cascallar, and Kyndt (2020); Sothan (2019).

Finally, the last group of determinants concerns academic performance parameters. They are classified as institution-related when concerning the place of study both in terms of infrastructures and programs offered, and student-related when regarding students' previous performances and curricula activities. Multiple examples can be found in the literature, for example in the works by Abdelfattah et al. (2022); Adeyemi and Adeyemi (2014); Alhajraf and Alasfour (2014); Danilowicz-Gösele et al. (2017); Gambini, Desimoni, and Ferretti (2022); Huang and Fang (2013); Torenbeek, Jansen, and Suhre (2013); Yousef (2019); Zuluaga-Ortiz, DelaHoz-Dominguez, and Camelo-Guarín (2022). The variable under investigation in this study, namely commuting time, falls within the category of institution-related factors, as it serves as a measure of university accessibility.

2.2. *Evidence of travel time in students' performance*

Recent studies have highlighted the potential impact of commuting time on students' cognitive functioning, stress, and overall well-being, and its effects are both positive and negative. On the positive side, commuting time provides individuals time to plan and organize their day, study, or review lessons, and can also serve as a mental break or a relaxing time when to listen to music or read a book. While it can be a time for personal reflection, learning, and development, excessive commuting time has significant drawbacks and negative consequences that cannot be ignored (Jachimowicz et al. 2021; Pindek, Shen, and Andel 2023; Wilhoit 2017). First of all, the advantages of travel time are limited in the case of high-quality commute, while are difficult to exploit in most situations, for instance, due to changes in means of transport or to noisy environments. Long and tiring commutes can be physically and mentally exhausting, and limit students in the opportunity to engage in sports or recreational activities, contributing to stress and anxiety (Costa, Pickup, and Di Martino 1988; Hansson et al. 2011; Künn-Nelen 2016). Given the ambivalent effect of travel time on well-being, it is important to dedicate further attention to in-depth investigation and analysis of this factor.

For the purposes of this research, an in-depth study is performed of those works that assess the relationship between the accessibility measure of commuting time to educational institutions with students' outcomes. Despite its significance, this topic remains relatively unexplored in the literature, particularly in higher education. It is crucial to delve deeper into this subject to better understand the potential consequences of long commuting times on academic achievement, and ultimately promote students' success in higher education. The examined studies are grouped based on the estimate used for

commuting time: geographical distance, estimate from surveys, and distance on the (weighted) road network.

The first measure of commuting time is the geographical distance. This approach assumes that the farther someone lives from their study institution, the longer their daily commute will be. The impact of distance between residence and college on educational attainment and GPA is investigated in the work by Garza and Fullerton (2018). They focus on a category of disadvantaged students, i.e. first-generation college students, and use a logistic regression model for the binary dependent variable of degree attainment and an Ordinary Least Squares (OLS) regression model for the GPA. Authors know information about ethnicity, age, gender, income, personal and family condition, academic skills, institutional characteristics, school selectivity, and social and academic integration. They find that a greater distance is related to a higher probability of obtaining a bachelor degree but they do not find evidence of a relationship between distance from home and students' GPA.

Despite its easy computation, a considerable drawback of the accessibility measure through the geographical distance between home and school institution is that it ignores factors such as traffic, transportation options, and the location of essential services, which can have a significant impact on the actual time spent commuting and are not neglectable in the context of metropolitan cities.

Another group of studies analyzes the information collected via questionnaires. This method is used in the works of Kobus, Van Ommeren, and Rietveld (2015) and Tigre, Sampaio, and Menezes (2017). The first work considers 2,857 students of the Vrije Universiteit in Amsterdam of which they know personal information, commuting time to university, year of attendance, degree program, and context information. The output variable considered is average grades and estimates are computed using OLS and Instrumental Variables (IV). Results of OLS show that the commuting time is not significant, while using IV the commuting time is shown to have a statistically significant negative impact on the outcome. The second research uses as output the grade obtained in a standardized mathematics test and collects data of sixth-grade students, their parents and school principals in 118 public schools in Brazil. The authors aim at estimate the impact of commuting time on students' performance using the distance to the closest two schools as IV, and they find strong evidence of a negative effect of time distance on performance.

Surveys are widely used in all fields of research, and many studies already highlighted the guidelines to avoid response bias. Still, literary works have established that people generally tend to underestimate their travel time, so for that reason, self-reported measurements of commuting time cannot be considered a reliable indicator of actual travel time (Stopher, FitzGerald, and Xu 2007).

The last group of works estimates the commuting time as the distance on the road network. Links of the road network are weighted by their expected travel time, usually related to their length, velocity constraints for pavement and street type, or more sophisticated time approximations. The goal of the work by Falch, Lujala, and Strøm (2013) is to investigate the impact of geographical proximity to upper secondary school students on graduation propensity using Norwegian data. They estimate the travel time between students' home ward midpoint and the nearest upper secondary school using ArcGIS Network Analysis. The distance is computed over the public road network and the travel is assumed by car at the speed limits. In the model, they control for personal characteristics, previous grades, socio-economic status, family context, and school environment. Also in this work, the results show a negative impact of travel time on school outcomes. Contreras Guajardo et al. (2018) study the causal effect of

commuting time over academic achievement on a sample of more than 23,000 eighth-grade students in Santiago. Student's performance is measured as the average score in Mathematics and Spanish on the Sistema de Medición de la Calidad de la Educación (SIMCE) test. Their analysis is performed in a multiple-step process: at the beginning, they compute the travel time from students' home and their university in three modes of transport (by car, by public transport, and by foot) using Google Maps API; then they predict the mode of transport used by the student by means of Machine Learning algorithms, including information about personal characteristics of students and home-school distance; finally, they use OLS, IV and Fixed Effect Models (FEM) to evaluate the effect of the estimated commuting time on the result of SIMCE test. In the work by Serra (2021), employing campus reallocation in Portugal as an exogenous variable, the impact of commuting on students' GPA is estimated by analyzing the changes in commuting times. The administrative dataset encompasses information on 5,546 students who pursued bachelor's degrees at Universidade Nova de Lisboa SBE between the academic years of 2007/08 and 2019/20. The results indicate a negative impact of commuting time on students' GPA.

These estimates are reliable indicators of commuting time as are able to include traffic conditions and potential delays on the expected travel route. However, they require the specification of the users' mean of transport or the users' average velocity to construct the estimate, so that different values are obtained for different means of transport and speed, and this assumption is a sharp limitation for the assessment of an overall travel time estimate.

To overcome the exposed limitations in commuting time estimation, the present research proposes the adoption of a new method that makes use of GPS smartphone data of anonymous individuals. GPS traces from smartphones are a relatively recent data source, spreading after the advent of new technologies. They are immensely valuable as they are automatically collected by smartphones effortlessly and at no apparent cost, simultaneously conveying insights about human mobility patterns. Research studies already exploited the valuable information of GPS data to study the human mobility patterns in metropolitan cities (Andrade, Cancela, and Gama 2020; Guan and Chen 2021; Joseph et al. 2020; Moro et al. 2021). When assessing the accessibility to healthcare services, mobility data sources such as GPS data have been used for the computation of accessibility maps to hospitals, clinics, and other healthcare facilities, like in Braga, Loureiro, and Pereira (2023) and Xia et al. (2019). To the authors' knowledge, this study is the first work in the education domain that makes use of GPS data for a data-driven commuting time estimation and analysis.

3. Methodology and Data

3.1. *From raw GPS data to commuting time estimation*

The aim of this step is to build commuting time estimates from any point of origin within the Milan municipality to the university campuses of Politecnico di Milano. These accessibility maps are then used to extract students' estimated commuting time based on their home residence location and their university campus.

Politecnico di Milano has two campuses in Milan: the Leonardo campus, located in a central area of the city, easily reachable by the urban subway and also by suburban and regional trains, and the Bovisa site, which lies in the north-west of the city and is connected by high-frequency suburban trains. Besides the geographical location, the

campuses differ by the programs of which they host face-to-face courses, namely each student attends classes in either one of the campuses depending on their program. The commuting time estimation phase took this peculiarity into account, estimating the accessibility maps of both campuses and then extracting each student's commuting time from the map of the sites they visit for classes.

The starting point for the construction of campus accessibility maps is the dataset of smartphone GPS data. The dataset has been provided by Cuebiq Inc., a specialized firm in location intelligence, collecting data in compliance with privacy regulations, and providing data for research purposes within their "Data for Good" program. The dataset consists of anonymized geo-localization records of mobile phones, containing the location information (latitude and longitude coordinates) at different time instances for Cuebiq users (provided via anonymized IDs). Additional information is also provided, such as the accuracy of the GPS measurement and the type of smartphone operating system that collects the data, as shown in Table 1. For the study, coordinates are transformed from the angular longitude and latitude into Universal Transverse Mercator (UTM) coordinates, which use a simpler Cartesian system in meters, to make distance calculations more straightforward compared to the angular measurements of latitude and longitude.

Table 1. GPS raw data description

Variable	Description	Type
Cuebiq ID	Anonymized device ID of Cuebiq user	Cat.
Device type	Mobile operation system (Android, iOS)	Cat.
Timestamp	Date and time in unix format	Num.
Longitude	Coordinate	Num.
Latitude	Coordinate	Num.
Accuracy	Accuracy of the measurement: radius [meters] of the circle around the captured GPS position with a 68% probability	Num.

To align with the project objectives, the original dataset is reduced to consider only observations in the period and area of interest. The selected observations are located in the Milan municipality and its surroundings in the period from December 1st to December 20th 2019, i.e. during one month of the first academic semester 2019/20, in working days (from Monday to Friday), and in morning peak hours (arrival between 7:30 and 9:30). The analysis of the reduced dataset allows to capture the human mobility patterns in the city of Milan that mimic the ones of students traveling to reach the university campus for attending the first lesson of the day. The analysis refers to a pre-COVID period, when all the lectures were in presence and students' attendance to face-to-face lectures was wide. The choice of this study period ensures to capture the actual relationship between students' performance and the duration of their home-to-university commute.

With the same goal, another necessary reduction consists in selecting the GPS traces of users whose journey ends at one of the two campuses of Politecnico di Milano, so that the final time estimation would be based on GPS data of actual travelers to Politecnico di Milano. A journey describes the spatial movements of a user between two periods of inactivity through a sequence of GPS data points ordered with respect to the timestamp. The identification of journeys from GPS raw data passes from the application of some filtering (e.g., remove low-accuracy data) and movement identification criteria (e.g., velocity above a certain threshold), which is described in detail in Appendix A. The resulting dataset for what concern the Leonardo campus is composed of 7,670 GPS measurements of 881 trajectories belonging to 448 distinct Cuebiq users.

Regarding the Bovisa campus, there are 4,111 GPS measurements of 498 trajectories belonging to 295 distinct Cuebiq users. Each of these GPS samples is associated with a measure of commuting time to the university campus, computed as the time elapsed from that measurement to the end of the journey (i.e., the arrival at the university).

To obtain the continuous distribution of commuting time from discrete smartphone GPS data, Kernel Regression Estimation (KRE) is applied. KRE is a non-parametric method to estimate the conditional mean of a response random variable Y conditioning on the values of other random variables $\boldsymbol{X}$. In the case of this analysis, the response variable Y represents the commuting time to the university campus and depends on the variables $\boldsymbol{X} = [X_1, X_2]$ describing the geographical coordinates. The KRE is performed by averaging the values of the response in the observed sample with weights given by kernel function evaluated at the covariate values. Specifically, given a sample of n observations of the response variable y_i and of the covariates $\boldsymbol{x}_i = [x_{1i}, x_{2i}]$, the kernel regression estimator of the conditional mean of Y in a new location $\boldsymbol{x}$ is expressed as:

$$\hat{m}(\boldsymbol{x}) = \hat{\mathbb{E}}[Y|\boldsymbol{X} = \boldsymbol{x}] = \frac{\sum_{i=1}^{n} K_H(\boldsymbol{x} - \boldsymbol{x}_i)\, y_i}{\sum_{i=1}^{n} K_H(\boldsymbol{x} - \boldsymbol{x}_i)}, \tag{1}$$

where K_H represents the bidimensional kernel function of bandwidth matrix H. The estimator in Equation 1 is the Nadaraya-Watson estimator (Nadaraya 1964; Watson 1964). Among the various kernel functions the one selected in this work is the Gaussian kernel, as suitable for this applicative problem and as the most commonly used in the literature.

In this work, the symmetry in the two spatial dimensions leads to the definition of the covariance matrix of the kernel as a positive scalar h times the 2×2 identity matrix, so that the same amount of smoothing is applied in the two coordinate directions. The value h is not held fixed but is varied depending on the location of the estimate and its neighbors, as a generalization of the Nearest-Neighbors (NN) regression. The generalized NN bandwidth is preferred in this work due to the inhomogeneity of the sample data distribution in the space. More formally, $h = h(\boldsymbol{x}) = c\, d_k(\boldsymbol{x})$, where c is a multiplying constant and d_k represents the Euclidean distance between the observation and its k-th nearest neighbor. The number of neighbors to be included in the bandwidth k and the multiplying constant c are chosen via the minimization of the Mean Square Error (MSE) of the Leave-One-Out Coss-Validation (LOO-CV) on a training set containing data of 85% trajectories. The remaining trajectory data are used as a test set to evaluate the performance of the method in terms of MSE.

KRE is applied to each of the two campuses of Politecnico di Milano, creating two accessibility maps that indicate the estimated commuting time to university from each point in the Milan municipality. A rectangular bounding box of extension 20.5×18.6km^2 is constructed to contain the whole area of the Milan municipality and an additional 1-kilometer wide buffer around it. Inside this box, a grid of 38,316 point locations is defined, with each point positioned 100 meters apart from its adjacent points. The KRE technique is then employed to estimate the commuting time to the university campus for these specified points.

Students' commuting times are determined by geolocating their addresses on the corresponding campus accessibility map and extracting the estimated times accordingly. Specifically, the point on the grid whose distance is the smallest to the residence location of the student is selected. The distance between these points is ensured to

be smaller than $50\sqrt{2}$ meters due to the point density of the grid, so the selected point can be considered a good proxy of the student's home location. The estimated commuting time computed for this location is then used as the student's estimated commuting time to university.

3.2. *Investigating the causal relationship between commuting time and academic performance*

The analysis focuses on first-year undergraduate students who were enrolled in engineering courses at Politecnico di Milano during the first semester of the academic year 2019/20. The sample is chosen primarily because the study plans exhibit similarity across all programs in this period of interest, then since first-year students are assumed to be more inclined towards attending classes, making them an ideal sample to analyze for the research.

Another important aspect of the study is its emphasis on students residing in Milan. This selection is motivated by two primary reasons. Firstly, focusing on students within this area enables the construction of accessibility maps with a high degree of precision. Secondly, the study assumes that students residing in Milan experience fewer significant life changes during their first year of university compared to students who relocated far away from their family and friends to pursue higher education, since the latter may face significant challenges that could influence their performance (Vieira, Vieira, and Raposo 2018). Additionally, the dataset is refined to consider only those students who passed at least one exam at the end of the first semester. This selection is necessary to exclude students for which the GPA cannot be computed due to the absence of exams taken. After the reductions, the dataset comprises 507 observations of first-year students in bachelor engineering programs who both reside in the Milan metropolitan area and passed at least one exam in the analyzed period. The Appendix B explores the representativeness of the selected sample.

The dataset incorporates various personal characteristics of each student, including gender, age of admission to the university, and family income. Additionally, information related to their educational background is considered, such as the final high school grade and the high school track. As for their current academic pursuits, the specific bachelor program followed is taken into account. This last element not only helps identify the campus attended by each student but also serves as a grouping criterion for students. Moreover, the commuting time to university, whose estimate is obtained as in Section 3.1, is considered. In terms of evaluating students' performance, output pertains to the GPA of students at the end of the first semester. The GPA is measured on the Italian grade scale, with marks from 1 to 30 and a threshold of sufficiency of 18.

To provide a clear overview of the variables involved and their characteristics, as well as to offer descriptive statistics, Table 2 summarizes all these factors.

The objective of the second step of the analysis is to look for the impact of students' commuting time on their students' performances. The analysis goes beyond the assessment of variable relationships and works in a causal inference domain. The investigation is on the cause-and-effect pattern linking the two variables of interest and representing the variations in the GPA that can be attributed to the variable of commuting time while controlling other factors (e.g., other students' covariates).

Formalizing the problem, let Y be the outcome variable, here representing the students' GPA. Let then A be the exposure variable, or continuous treatment variable,

Table 2. Students' variables description and descriptive statistics

Variable's name	Variable's Description	Typology	Descriptive statistics
Gender	Student's gender	Categorical	Male: 370 (73%) [base] Female: 137 (27%)
Admission age	Student's age at the admission to university (baseline 18)	Numeric	[-1;14] mean=0.57 SD=1.11
Family income	Student's family income	Categorical	High: 281 (55%) [base] Middle: 96 (19%) Low: 38 (8%) Grant: 92 (19%)
High school grade	Standardized high school final grade	Numeric	[0.6-1] mean=0.82 SD=0.11
High school track	High school track attended by the student	Categorical	Humanistic: 29 (6%) Scientific: 444 (88%) [base] Technical: 26 (5%) Other: 8 (2%)
Commuting time	Estimated commuting time in the first step expressed in hours	Numeric	[0-0.74] mean=0.38 SD=0.17
GPA	GPA of the student at the end of the first semester of 2019/20 (baseline 18)	Numeric	[0-12] mean=5.6 SD=3.04

here describing the students' commuting time, and define the vector of additional covariates of the dataset $\boldsymbol{X}$. The interest is in knowing the effect of A on the outcome Y controlling over $\boldsymbol{X}$. This is done by looking at the estimated expected value of the outcome Y for different values a of the exposure variable A, that is, by looking at the estimate of the function $\mu(a) = \mathbb{E}[Y(a)]$.

The analyzed dataset is observational, meaning it collects measurements of subjects through the external observation of phenomena. Unlike randomized experiments, where treatment can be randomly assigned to subjects, observational studies record data without manipulation and control over them. As a consequence, it is not guaranteed that the relationship that emerges between the outcome and treatment is not affected by other factors that act as confounders. In the case study, for instance, students' characteristics may have an imbalanced distribution between different levels of commuting time, affecting the evaluation of the relationship between the outcome GPA and the commuting time and, hence, working as confounders. Therefore, measuring the impact of the exposure on the outcome is not immediate and comes as a challenge. The Potential Outcome Framework (Neyman 1938; Rubin 2005) addresses this challenge by formalizing the concept of counterfactuals and offering a systematic way to estimate causal effects in dataset. Within this framework, the Propensity Score and its generalization to the continuous treatment case are the starting point of many methods for observational studies that naturally deal with confounding (Hirano and Imbens 2004; Rosenbaum and Rubin 1983). The Generalized Propensity Score summarizes the information about potential confounders into a single score by measuring how likely the combination of exposure and characteristics of a sample unit is represented in the analyzed sample. To formalize its definition, let r denote the conditional density of the continuous treatment variable given the covariates, $r = r(a, \boldsymbol{x}) = f_{A|\boldsymbol{X}}(a|\boldsymbol{x})$.

Then, the Generalized Propensity Score is $R = r(A, \boldsymbol{X})$.

Among the four classes of Propensity Score techniques used to remove the effects of confounding outlined by Austin (2011), balancing weighting methods are the most popular for dealing with the case of continuous exposure, moving forward from the classic case of binary treatment. These methods work by assigning a weight to each observation in the dataset, measuring the importance that the instance needs to have in the analyzed dataset to make it fully balanced: larger weights are given to subjects who are less common in the study population, ensuring their influence is appropriately accounted for. The state-of-the-art techniques for the computation of balancing weights work by estimating the Generalized Propensity Score and using its inversion as a weight. However, in the last ten years, the goal of covariate balancing has been addressed with other innovative methodologies that either integrate the Generalized Propensity Score estimation in wider covariate balancing methods or assess the dataset balance by looking only at the covariate distribution.

The present research follows the newest works in the field of balancing methods addressing the impact of continuous treatment on the outcome variable. Different weighting methods are explored and compared to find the most suitable for the application. Subsection 3.2.1 is dedicated to the description of such weight comparison and application. Then, the selected weights are applied to the analyzed sample and the impact of commuting time on students' GPA is investigated, as discussed in Subsection 3.2.2.

3.2.1. Dataset balancing

The goal of this subsection is to remove the effect of external confounding factors. Apart from the commuting time and student's GPA, which here work as continuous exposure A and outcome Y respectively, all the observed covariates reported in Table 2 are included as confounders. It is assumed that no other unobserved determinants of students' performances confound the relationship, since the available covariates already describe students' characteristics from the point of view of the three classes of determinants explained in Section 2.

For the selection of the proper methodology, four weighting methods already available in the literature are compared with respect to their balancing performance in this application case. Specifically, this research employs and compares the Inverse Probability Weights computed from the Generalized Propensity Score, the application of Covariate Balancing Propensity Score, both in the parametric and non-parametric case, and the weight estimation from the Entropy Balancing method.

In the first approach, following the path of Rosenbaum and Rubin (1983) and the subsequent generalization of Hirano and Imbens (2004) which deals with the case of continuous exposure, Inverse Probability Weights (IPWs) are computed from the inversion of Generalized Propensity Scores of the observed data. The IPWs are obtained in such a way that higher weights are associated with less-represented sample units and vice-versa, balancing the distribution of confounding characteristics across exposure levels.

The second weighting method analyzed is the Covariate Balancing Propensity Score (CBPS). Proposed by Imai and Ratkovic (2014), the method parallelly estimates the balancing weights and models the relationship between exposure and confounders. It is able to overcome the limitation of the classic IPWs of suffering for substantial bias in the estimation after slight misspecification of the Generalized Propensity Score model. Both the parametric and non-parametric versions of the CBPS are used and compared

for weight computation (Fong, Hazlett, and Imai 2018).

Finally, the Entropy Balancing (EB) weighting method is considered (Hainmueller 2012; Vegetabile et al. 2021). Differently from previously mentioned methods, EB weights are computed without a direct estimation of the generalized propensity score, but through the minimization of a measure of the dispersion of the weights, that is the entropy, while controlling conditions on specified moments of the covariates. The method removes the requirement of modeling the generalized propensity score and shows performance comparable to the one of CBPS.

Most research works on datasets of independent observations with a non-hierarchical structure, for which the theoretical and practical framework of generalized propensity score and balancing methods is already developed and consolidated. However, in this research application, data have a hierarchical structure (i.e., students belong to different bachelor programs). As pointed out by many as Arpino and Mealli (2011) and Li, Zaslavsky, and Landrum (2013), the clustered nature of the data is a source of relevant information and should not be discarded but integrated into the weighting method. For the case of continuous exposure, Schuler, Chu, and Coffman (2016) propose to include the grouping factor in the model for generalized propensity score estimation. They use both linear Fixed Effects Models (FEMs) and Random Effects Models (REMs) for modeling the exposure variable from the confounders and getting an estimate of the generalized propensity score and of the IPWs.

Starting from these studies, the current analysis includes the bachelor program information so as not to neglect the hierarchical structure of the data. Previously mentioned methods are modified: for the computation of IPWs, the program is added either as a fixed effect or as a random effect in the formulation of the generalized propensity score (i.e., through FEMs and REMs); for the CBPS methods, the program is added as a fixed effect in the formulation of the generalized propensity score as in FEMs; likewise, for the EB method, dummy covariates of the bachelor program are added as confounders, in accordance with the FEM approach.

More specifically for what concerns the IPW computation, the model formulation for generalized propensity score estimation is expressed in Model 2 for FEM and Model 3 for REM.

$$t_{il} = \gamma_0 + \sum_{j=1}^{J-1} \gamma_j x_{jil} + \sum_{l=1}^{L} z_l + \epsilon_{il}, \quad \epsilon_{il} \sim \mathcal{N}(0, \sigma_{\epsilon_1}^2) \tag{2}$$

$$t_{il} = \gamma_0 + \sum_{j=1}^{J-1} \gamma_j x_{jil} + z_l + \epsilon_{il}, \quad z_l \sim \mathcal{N}(0, \sigma_z^2), \quad \epsilon_{il} \sim \mathcal{N}(0, \sigma_{\epsilon_2}^2) \tag{3}$$

In the formulas 2 and 3, t_{il} indicates the commuting time of student $i \in \{1, \dots, n_l\}$, $n = \sum_l n_l$, in program $l \in \{1, \dots, L\}$; $\boldsymbol{\gamma} = \{\gamma_0, \dots, \gamma_{J-1}\}$ is the $J-$dimensional vector of parameters; x_{jil} represents the value of the $j-$th confounding student covariate; z_l is the either fixed (in FEM) or random (in REM) intercept of the program l; ϵ_{il} is the error; in REM it is assumed that $\boldsymbol{z}$ is independent of $\boldsymbol{\epsilon}$.

After their computation, weights are compared following the guidelines proposed in the literature (Vegetabile et al. 2021; Zhu, Coffman, and Ghosh 2015). The comparison is made by controlling the following quantities:

- Correlation balance: it involves the evaluation of the Pearson correlation between the exposure variable and the entire set of observed covariates;
- Marginal balance: authors examine marginal summary statistics for each variable in the dataset, including both the covariates and the exposure variable. This analysis considers both mean and variance measurements;
- Effective Sample Size: this summary quantifies the effective number of observations within each sample.

By investigating these summaries in both weighted and unweighted distributions, authors can comprehensively assess the performance of the different weighting methods and choose the optimal one for the application. Results of this analysis are shown in Subsection 4.2

3.2.2. Modeling the impact

Once the optimal weights are selected, they are used to balance the original dataset and train a regression model on it. Through these analyses, the aim is to achieve a holistic understanding of the influence of commuting time on students' performance in a broader perspective, beyond the correlation between these two variables. Additionally, such correlation is investigated with the application of the same regression model to the original (unbalanced) dataset, allowing both to compare the results from the two training datasets and also to get proper insight into the relationship between GPA and other students' factors.

The regression model chosen for modeling the outcome is a Linear Mixed-effect Model (LMM) (Goldstein 2011). This model allows for the analysis of individual student data while accounting for the variation within the different grouping structures, i.e. the bachelor programs. This is essential as different programs may have distinct characteristics and requirements that can affect students' academic performance (e.g., peer effect, teacher effect). Moreover, to gain deeper insights into the impact of students' commuting time on their GPA, a Polynomial Regression analysis is conducted. While simple linear models are known for their simplicity in description and implementation, along with their advantages in terms of interpretation and inference, polynomial linear models take this approach a step further. They expand simple linear models by introducing additional predictors, achieved by raising the original predictors to various powers so that they allow the description of nonlinear relationships between covariates and response (James et al. 2013). Specifically, for this study, a model with commuting time raised to the Dth-power is chosen, since it has been identified as the optimal power leading to the lowest MSE when applied to a model trained on 70% of the original observations and subsequently evaluated on the remaining.

All the available covariates are included in the model: GPA as the response variable, and commuting time and other determinants as predictors. When using the weighted dataset, although it has already been balanced with respect to other covariates than the commuting time, the latter are still included to reduce the bias due to residual imbalance and to gain robustness in the effect estimate (Ho et al. 2007). Nevertheless, when employing the weighted dataset, the estimates of their coefficients are not interpreted nor examined since may be severely confounded and the final considerations regard the effect of the commuting time only. While the bachelor program is included in the model as a random effect, the remaining determinants are included in the model as fixed effects. Numerical variables appear with an additive relationship to the response. Categorical and binary variables are introduced as dummy variables with respect to a defined baseline, again additive on the response.

13

The proposed model can be written as:

$$y_{il} = \beta_0 + \sum_{j=1}^{J-1} \beta_j x_{jil} + \sum_{d=1}^{D} \beta_{(J-1+d)} a_{il}^d + u_l + \epsilon_{il}$$

$$u_l \sim \mathcal{N}\left(0, \sigma_u^2\right)$$

$$\epsilon_{il} \sim \mathcal{N}\left(0, \sigma_\epsilon^2\right)$$

(4)

where y_{il} denotes the grade point average of student i in program l; $\boldsymbol{\beta} = \{\beta_0, \ldots, \beta_{J+D-1}\}$ is the $(J + D)-$dimensional vector of parameters; x_{jil} represents the value of the $j-$th predictor at student's level, in particular a_{il} is the variable of commuting time up to degree D; u_l is the random effect of the program l; ϵ_{il} is the error; and it is assumed that $\boldsymbol{u}$ is independent of $\boldsymbol{\epsilon}$.

The effect of commuting time on GPA is estimated by looking at the Average Dose-Response Function (ADRF) and the Average Marginal Effect Function (AMEF). The first function links the value of the commuting time to the expected potential GPA under that travel time value across the full sample and estimates the average potential outcome across different levels of exposure: $AMEF(a) = \hat{\mu}(a) = \hat{\mathbb{E}}[Y(a)]$. 60 distinct values of commuting time are examined, ranging from its 5%-quantile to its 95%-quantile, and the estimate is computed together with its 90% confidence intervals. AMEF is also used to characterize the effect of commuting time with the derivative of the ADRF with respect to travel time. When different from zero, it highlights a significant impact of commuting time for its corresponding values. The two estimated curves are compared for the cases of unweighted and weighted datasets, showing the relevance of the balancing procedure for the current application.

4. Results

4.1. Commuting time estimation

The commuting time estimation procedure is applied separately to the two campuses of the Politecnico di Milano. The $N_L = 7,670$ input GPS measurements concerning the Leonardo campus are depicted in Figure 2a, while the $N_B = 4,111$ data for the Bovisa site are shown in Figure 2b. They both span within the rectangular box that entirely contains the municipality of Milan, indicated by a gray outline in the figures. The figures also show the location of the university institutions, colored in blue, and the main public transportation lines, with suburban train lines colored in green and subway lines colored in red.

The choice of the hyperparameters k and c for KRE is performed through LOO-CV on the training set. As shown in Figure 3, the MSE is minimized by choosing the combination of $k = 0.005N_L$ and $c = 1/3$ for the Leonardo site (MSE = 32.62 min^2), and $k = 0.01N_B$ and $c = 1/3$ for the Bovisa site (MSE = 44.78 min^2). With such parameters, the performance of the methodology is evaluated on a test set of 15% of the trajectory data. The KRE of commuting time to the Leonardo campus results to be highly accurate, with MSE = 22.43 min^2 and Mean Absolute Error (MAE) = 3.24 min. Concerning the Bovisa site, the performance slightly decreases with MSE = 41.83 min^2 and MAE = 4.62 min.

The estimated commuting time for each point location on the grid is shown in Figure

(a) Leonardo campus

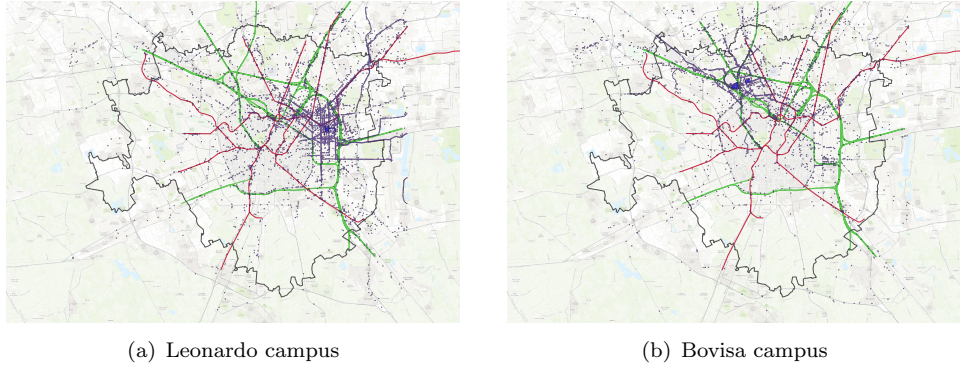(b) Bovisa campus

Figure 2. Input GPS data within the analyzed area of Milan municipality for the two campus sites.
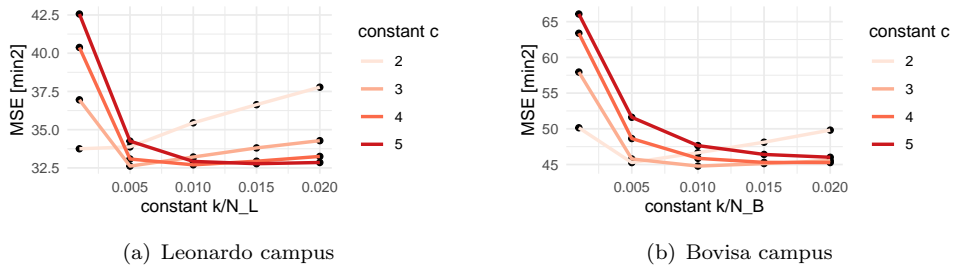

(a) Leonardo campus

(b) Bovisa campus

Figure 3. MSE of the LOO-CV performed for different parameters k and c.

4 for the accessibility to the Leonardo campus (a) and to the Bovisa campus (b). For both maps, the estimated times generally increase as the distance to the institution increases, as expected. However, it is interesting to point out that the increase does not occur homogeneously in space, but there are directions of anisotropy that follow the main lines of the road network and the public transportation network. For instance, for the Leonardo case, locations in the southwest of the analyzed box, although at a high distance to the university campus, show a low estimated travel time to university when near the subway network. Similarly, for the Bovisa campus, the presence of the highway in the northwest and of the suburban train line in the southeast decreases the estimated commuting time to Bovisa with respect to other locations at the same distance from the university but far from the transportation network.
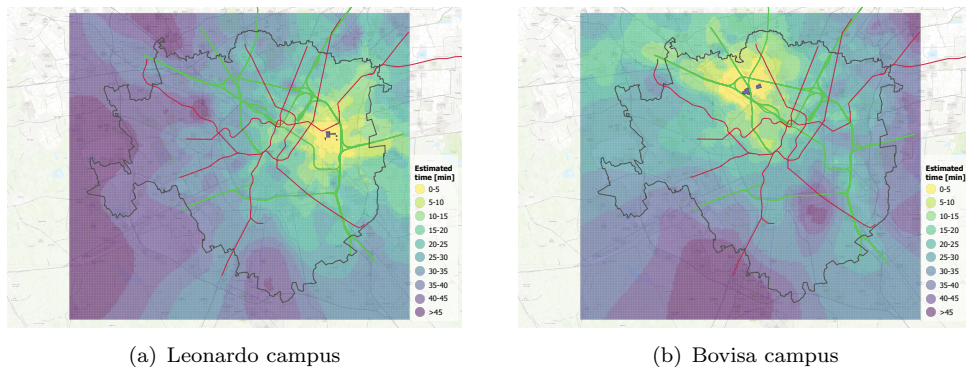

(a) Leonardo campus

(b) Bovisa campus

Figure 4. Accessibility maps of estimated commuting time, output of KRE, for the two university campuses.

15

Table 3. Weights comparison: correlation exposure - confounders

	NO	npCBPS	CBPS	GLM	EB	REM	FEM	GLM_{ml}	EB_{ml}	$npCBPS_{ml}$	$CBPS_{ml}$
AdmissionAge	0.030	0.00069	0.0039	0.015	0.0017	0.018	0.019	0.017	0.0061	0.0032	0.00069
Gender	-0.073	0.0084	0.0038	0.0046	0.0040	0.0090	0.0098	0.015	0.00087	0.0067	0.00057
HighSchoolTrack	0.044	0.0097	0.011	0.0063	0.011	0.0055	0.0017	-0.0025	0.010	0.0062	0.0096
HighSchoolGrade	-0.054	-0.0095	-0.0029	0.0018	-0.0030	-0.0057	-0.016	-0.017	0.000055	0.0040	0.0024
FamilyIncome	0.039	-0.00095	0.0030	0.015	0.0027	0.020	0.024	0.026	0.00029	-0.0032	0.00040
Bachelorprogram	-0.016	-0.011	-0.0083	-0.0085	-0.0091	-0.011	-0.016	-0.016	-0.0053	-0.0055	-0.0054

Table 4. Weights comparison: descriptive statistics.
Descriptive statistics of the balanced samples compared with those of the original sample size. The mean value and standard deviation are reported for continuous variables while proportion per group for categorical ones

	NO	npCBPS	CBPS	GLM	EB	REM	FEM	GLM_{ml}	EB_{ml}	$npCBPS_{ml}$	$CBPS_{ml}$
AdmissionAge	0,566	0,564	0,577	0,564	0,566	0,569	0,573	0,570	0,566	0,566	0,592
	(1,106)	(1,122)	(1,190)	(1,113)	(1,124)	(1,112)	(1,105)	(1,095)	(1,090)	(1,080)	(1,206)
Gender											
F	27.0%	27.0%	26.8%	26.7%	27.0%	27.0%	27.3%	27.4%	27.0%	27.0%	26.4%
HighSchoolTrack											
Humanistic	5.72%	5.72%	5.57%	5.56%	5.72%	5.62%	5.75%	5.71%	5.72%	5.73%	5.55%
Scientific	87.57%	87.57%	87.63%	87.70%	87.58%	87.70%	87.76%	87.87%	87.57%	87.56%	88.01%
Technical	5.13%	5.14%	5.14%	5.16%	5.13%	5.11%	4.99%	4.94%	5.13%	5.13%	4.83%
HighSchoolGrade	0,822	0,822	0,823	0,824	0,822	0,824	0,824	0,825	0,822	0,822	0,823
	(0,110)	(0,111)	(0,111)	(0,111)	(0,111)	(0,111)	(0,112)	(0,112)	(0,111)	(0,111)	(0,110)
FamilyIncome											
high	55.42%	55.44%	55.28%	55.25%	55.43%	55.03%	55.19%	55.11%	55.42%	55.44%	55.03%
medium	18.93%	18.90%	19.18	19.28	18.93%	19.41%	19.29%	19.41%	18.93%	18.92%	18.87%
low	7.50%	7.50%	7.62%	7.51%	7.49%	7.47%	7.44%	7.41%	7.50%	7.50%	7.44%

Table 5. Weights comparison: effective sample size.

	NO	npCBPS	CBPS	GLM	EB	REM	FEM	GLM_{ml}	EB_{ml}	$npCBPS_{ml}$	$CBPS_{ml}$
ESS	507	492	494	494	495	488	471	465	456	432	405

4.2. *Commuting time and GPA: relationship and impact*

The optimal balancing technique for the study application is identified after comparison with the aim of selecting the best balance for the dataset. Table 3, Table 4, and Table 5 present the analyzed metrics for balance comparison: correlation balance, marginal balance, and effective sample size, computed for the original dataset and the datasets balanced employing weighting methods.

In the absence of any balancing techniques, the correlation values between travel time and confounding variables span a range from -0.073 to 0.044. This observation underscores the inherent balance within the original dataset. Upon evaluating various balancing approaches, the most significant reduction in correlation is achieved when implementing the methods of EB, npCBPS, and CBPS with multilevel correction.

Each of the employed balancing methods rigorously preserves the integrity of descriptive statistics within the balanced dataset, ensuring that they remain unaltered when compared to their counterparts in the original dataset. This robust consistency reaffirms the reliability and effectiveness of the balancing techniques applied.

The integrity of the balanced samples is consistently upheld, with their effective size consistently remaining at or above 80% of the original sample size. It's noteworthy, however, that the application of methods incorporating multilevel correction does lead to slightly smaller effective sample sizes compared to other methods.

Based on these insightful findings, authors are able to make an informed selection of the most suitable balancing method. It's crucial to highlight that all the balancing methods employed can be deemed appropriate for the specific objectives of the project. For the upcoming analyses, authors have opted to implement the Entropy Balancing method with multilevel correction, leveraging its demonstrated effectiveness in optimizing the dataset for further investigation.

Finally, the relationship and the impact of commuting time on GPA are investigated through the multilevel model in Model 4. The model is firstly trained on the original unweighted dataset and secondly on the balanced weighted dataset. Analyzing the unweighted dataset allows for both delving into the determinants influencing students' school performance and for obtaining baseline results on the time-GPA relationship, to be further compared with the time-GPA impact. Parallelly, the analysis of the balanced dataset leads to the identification of the causal impact of interest.

Regarding the unweighted dataset, at first, the optimal degree for the variable representing students' commuting time estimate is determined via MSE minimization. Figure 5 shows the trend of the MSE on the test set when the polynomial in the model is defined for different degrees $D \in \{1, \ldots, 10\}$ on a training set of 70% individuals. The model that minimizes the MSE is the one of degree $D = 5$.

The performance of the model is evaluated through LOO-CV on the complete dataset and returns a Root Mean Square error (RMSE) of 2.785. The significant covariates are the gender of the student, with a negative impact of -0.66 for being a female, the previous achievements in high school, increasing the average GPA of 0.1109 for every point in the high school grade, and the high school track, penalizing students from humanistic high schools than scientific ones. Figure 6 collects these results. Interestingly, the commuting time remains significant for explaining the variability of the GPA. The AMEF of commuting time on the model outcome, shown in Figure 7 with a grey line, is significantly different from zero when the commuting time is less than 15 minutes and for higher values between 30 and 40 minutes, while there is no impact of commuting time for journeys of medium duration between 15 minutes and 30 minutes.

The marginal effect of commuting time on students' performances is computed for the balanced weighted dataset and analyzed. Both from the ADRC and the AMEF, shown in Figure 7 with an orange color, the impact of commuting times between 5 and almost 15 minutes is significant on students' performances at the 90% confidence level, almost-linearly decreasing the average grade from 25/30 to 23.5/30. For larger times, however, travel time does not have a statistically significant effect on academic performance, being the AMEF non-significantly different from zero for travel times larger than 15 minutes. These results partially align with the ones of the analysis of the unbalanced dataset, showing the same general behavior of the curves, as in Figure 7 colored in grey. It is interesting to notice, however, that the marginal effect of commuting time would have resulted significant at 90% also for the range of values around 35 minutes that, thanks to the balancing procedure, is contrarily spotted to have a non-significant effect.
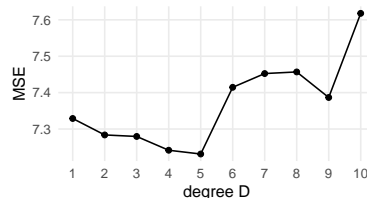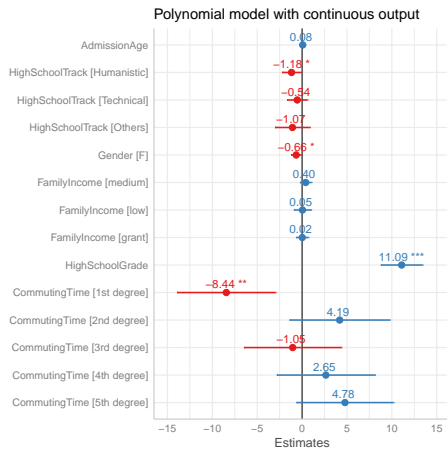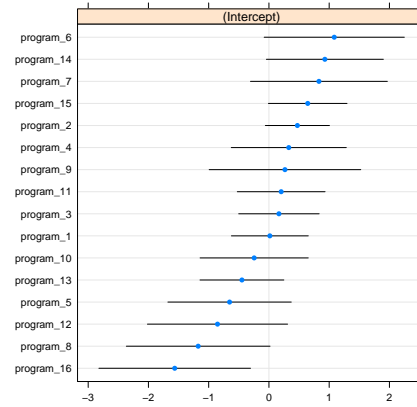

Figure 5. MSE (y-axis) evaluated on the test set when models of different degrees (x-axis) are trained.
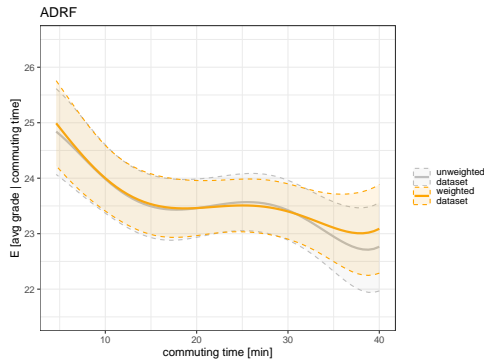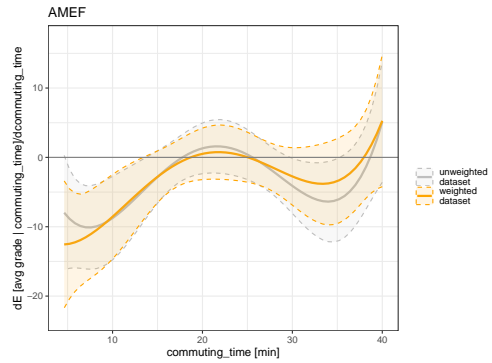
17

(a) Estimated fixed coefficients

(b) Random effects

Figure 6. Estimated coefficients of the polynomial model with continuous output (Mod. 4).

(a) Average Dose-Response Function

(b) Average Marginal-Effect Function

Figure 7. (a) ADRF of commuting time on the potential outcome, with confidence bands of 90%; (b) AMEF relating commuting time to the derivative of the ADRF, with confidence bands of 90%. In the case of the unbalanced dataset lines are shown in grey, while for the balanced dataset through Entropy weights they are depicted in orange.

5. Conclusions

This study implies a two-step procedure to evaluate the effect that commuting time has on the academic performance of university students focusing on the metropolitan area of Milan and the students of the Politecnico di Milano. The first step of the analysis uses machine learning algorithms and GPS data from anonymous users to reconstruct accessibility maps of the two university campuses. The latter leads to the estimation of commuting time for each student, whose information is used in the next step. The second step involves an analysis of the determinants of student performance in terms of GPA at the end of the first semester. This analysis reveals a negative dependency with statistical significance of travel time on student performance, which is why this study aims to go deeper and investigate the causal relationship of these factors. A search for a suitable balancing method follows and the selected one turns out to be the Entropy Balancing method with multilevel correction. Finally, in the causal inference setting the previous conclusions are confirmed.

The innovation and importance of this work lie in the application of the current analytical and methodological framework, but also in the implications of this study. It aims to analyze a key determinant of students' academic performance, i.e. commuting time, that has not been extensively and thoroughly investigated in the literature to date. In the first phase of the study, an innovative data source such as GPS measurements is used and a new methodology specific to this context is adopted. The results show that the proposed data-driven approach is crucial to accurately represent the mobility dynamics of a metropolitan city like Milan, where physical distance is not a reliable measure of accessibility, and where the transportation network is influenced by service availability and traffic conditions. In the second phase of the study, on the other hand, a causal relationship is examined, using balancing methods drawn from the recent literature and once again adopting an innovative methodological approach for this field. To the best of the authors' knowledge, this is the first time that the causal relationship between commuting time and academic performance is highlighted and statistically tested using recent techniques of causal inference like balancing weighting methods.

The implications of this research are highly relevant for the improvement and support of university students. One of the key findings underscores the crucial importance of positioning university campuses in well-connected and easily accessible areas within the city. This discovery suggests that, when planning new campuses, prioritizing strategic locations to ensure easy accessibility for students from various parts of the city should be paramount. This initial step can significantly enhance the learning experience of students and contribute to optimize their academic performance. Simultaneously, it is essential to develop and support an efficient and well-structured urban transportation system designed to make commuting to existing campuses as seamless as possible. The goal is to reduce commuting times for students, allowing them to allocate more time to their studies and academic activities. Another significant implication of this study is the need to strategically position student residences in relation to the campuses. If housing facilities are situated in close proximity to the study locations, students will benefit from substantial reductions in commuting times to attend classes. This not only enhances the quality of students' lives but can also contribute to increased participation in academic activities and engagement in university life. Finally, it is important to highlight that the findings of this study may have relevance in urban contexts other than the city of Milan. The methodology used and the implications drawn could be replicated and adapted to other cities, thereby

contributing to the optimization of the learning environment for university students in settings similar to the Milan municipality. In line with this, future developments could focus on assessing the reproducibility of this study in similar contexts, encompassing not only other Italian metropolitan cities but also European ones. Moreover, to enhance its value and appeal, it would be highly advantageous to broaden the scope of analysis beyond the central area of Milan. This entails considering the surrounding areas (e.g., the province of Milan) and incorporating the perspective of commuting students.

Acknowledgement and funding

References

Abdelfattah, Faisal A., Omar S. Obeidat, Yousef A. Salahat, Maha B. BinBakr, and Adam A. Al Sultan. 2022. "The predictive validity of entrance scores and short-term performance for long-term success in engineering education." *Journal of Applied Research in Higher Education* 14 (4): 1272–1285.

Abu Saa, Amjed, Mostafa Al-Emran, and Khaled Shaalan. 2019. "Factors affecting students' performance in higher education: a systematic review of predictive data mining techniques." *Technology, Knowledge and Learning* 24: 567–598.

Adeyemi, Abisola Moradeyo, and Seminu Babatunde Adeyemi. 2014. "Institutional Factors as Predictors of Students' Academic Achievement in Colleges of Education in South Western Nigeria." *International Journal of Educational Administration and Policy Studies* 6 (8): 141–153.

Alhajraf, Nout M., and Aishah M. Alasfour. 2014. "The impact of demographic and academic characteristics on academic performance." *International Business Research* 7 (4): 92–100.

Andrade, Thiago, Brais Cancela, and João Gama. 2020. "From mobility data to habits and common pathways." *Expert Systems* 37 (6): e12627.

Arel-Bundock, Vincent. 2023. *marginaleffects: Predictions, Comparisons, Slopes, Marginal Means, and Hypothesis Tests*. R package version 0.12.0, `https://CRAN.R-project.org/package=marginaleffects`.

Arpino, Bruno, and Fabrizia Mealli. 2011. "The specification of the propensity score in multi-level observational studies." *Computational Statistics & Data Analysis* 55 (4): 1770–1780.

Austin, Peter C. 2011. "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies." *Multivariate Behavioral Research* 46 (3): 399–424.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67 (1): 1–48.

Braga, Carlos Kaue V., Carlos Felipe Grangeiro Loureiro, and Rafael H. M. Pereira. 2023. "Evaluating the impact of public transport travel time inaccuracy and variability on socio-spatial inequalities in accessibility." *Journal of Transport Geography* 109: 103590.

Contreras Guajardo, Dante, Daniel Hojman Trujillo, Manuel Matas, Patricio Rodríguez Perales, and Nicolás Suárez. 2018. *The impact of commuting time over educational achievement: A machine learning approach*. Technical Report. Universidad de Chile. Facultad de Economía y Negocios.

Corazzini, Luca, Silvia D'Arrigo, Emanuele Millemaci, and Pietro Navarra. 2021. "The influence of personality traits on university performance: Evidence from Italian freshmen students." *PLoS One* 16 (11): e0258586.

Costa, Giovanni, Laurie Pickup, and Vittorio Di Martino. 1988. "Commuting–a further stress factor for working people: evidence from the European Community. II. An empirical study." *International archives of occupational and environmental health* 60 (5): 377–385.

Danilowicz-Gösele, Kamila, Katharina Lerche, Johannes Meya, and Robert Schwager. 2017.

"Determinants of students' success at university." *Education economics* 25 (5): 513–532.

Eurostat. 2020. "Majority commuted less than 30 minutes in 2019." `https://ec.europa.eu/eurostat/web/products-eurostat-news/-/ddn-20201021-2`.

Falch, Torberg, Päivi Lujala, and Bjarne Strøm. 2013. "Geographical constraints and educational attainment." *Regional Science and Urban Economics* 43 (1): 164–176.

Flom, Jacalyn M., Karen Y. Green, and Steven Wallace. 2021. "Helping higher education students succeed: An examination of student attributes and academic grade performance." *Active Learning in Higher Education* 24 (2): 221–235.

Fong, Christian, Chad Hazlett, and Kosuke Imai. 2018. "Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements." *The Annals of Applied Statistics* 12 (1): 156–177.

Gambini, Alessandro, Marta Desimoni, and Federica Ferretti. 2022. "Predictive tools for university performance: an explorative study." *International Journal of Mathematical Education in Science and Technology* 1–27.

Garza, Alma Nidia, and Andrew S. Fullerton. 2018. "Staying close or going away: How distance to college impacts the educational attainment and academic performance of first-generation college students." *Sociological Perspectives* 61 (1): 164–185.

Geurs, Karst T., and Bert van Wee. 2004. "Accessibility evaluation of land-use and transport strategies: review and research directions." *Journal of Transport Geography* 12 (2): 127–140.

Goldstein, Harvey. 2011. *Multilevel statistical models.* John Wiley & Sons.

Greifer, Noah. 2023. *WeightIt: Weighting for Covariate Balance in Observational Studies.* R package version 0.14.1, `https://CRAN.R-project.org/package=WeightIt`.

Guan, Xiangyang, and Cynthia Chen. 2021. "A behaviorally-integrated individual-level state-transition model that can predict rapid changes in evacuation demand days earlier." *Transportation Research Part E: Logistics and Transportation Review* 152: 102381.

Gwosc, Christoph, Kristina Hauschildt, Froukje Wartenbergh-Cras, and Hendrik Schirmer. 2021. *Social and Economic Conditions of Student Life in Europe: Eurostudent VII 2018-2021 – Synopsis of Indicators.* wbv Media GmbH & Company KG.

Hainmueller, Jens. 2012. "Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies." *Political analysis* 20 (1): 25–46.

Hansson, Erik, Kristoffer Mattisson, Jonas Björk, Per-Olof Östergren, and Kristina Jakobsson. 2011. "Relationship between commuting and health outcomes in a cross-sectional population survey in southern Sweden." *BMC public health* 11: 1–14.

Hanushek, Eric A. 1979. "Conceptual and empirical issues in the estimation of educational production functions." *Journal of human Resources* 14 (1): 351–388.

Hayfield, Tristen, and Jeffrey S. Racine. 2008. "Nonparametric Econometrics: The np Package." *Journal of Statistical Software* 27 (5): 1–32. `https://www.jstatsoft.org/index.php/jss/article/view/v027i05`.

Hellas, Arto, Petri Ihantola, Andrew Petersen, Vangel V Ajanovski, Mirela Gutica, Timo Hynninen, Antti Knutas, Juho Leinonen, Chris Messom, and Soohyun Nam Liao. 2018. "Predicting academic performance: a systematic literature review." In *Proceedings companion of the 23rd annual ACM conference on innovation and technology in computer science education*, 175–199.

Hirano, Keisuke, and Guido W. Imbens. 2004. *The Propensity Score with Continuous Treatments*, Chap. 7, 73–84. John Wiley Sons, Ltd.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15 (3): 199–236.

Hu, Shan, Rongtian Zhao, Yuxue Cui, Die Zhang, and Yong Ge. 2023. "Identifying the uneven distribution of health and education services in China using open geospatial data." *Geography and Sustainability* 4 (2): 91–99.

Hu, Xiao, Christy WL Cheong, Wenwen Ding, and Michelle Woo. 2017. "A systematic review of studies on predicting student learning outcomes using learning analytics." In *Proceedings*

of the seventh international learning analytics & knowledge conference, 528–529.

Huang, Shaobo, and Ning Fang. 2013. "Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models." *Computers Education* 61: 133–145.

Imai, Kosuke, and Marc Ratkovic. 2014. "Covariate balancing propensity score." *Journal of the Royal Statistical Society: Series B: Statistical Methodology* 76 (1): 243–263.

Jachimowicz, Jon M., Julia Lee Cunningham, Bradley R. Staats, Francesca Gino, and Jochen I. Menges. 2021. "Between home and work: Commuting as an opportunity for role transitions." *Organization Science* 32 (1): 64–85.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An introduction to statistical learning*. Vol. 112. Springer.

Joseph, Lucy, An Neven, Karel Martens, Opportuna Kweka, Geert Wets, and Davy Janssens. 2020. "Measuring individuals' travel behaviour by use of a GPS-based smartphone application in Dar es Salaam, Tanzania." *Journal of Transport Geography* 88: 102477.

Kobus, Martijn B.W., Jos N. Van Ommeren, and Piet Rietveld. 2015. "Student commute time, university presence and academic achievement." *Regional Science and Urban Economics* 52: 129–140.

Kumar, Mukesh, AJ Singh, and Disha Handa. 2017. "Literature survey on student's performance prediction in education using data mining techniques." *International Journal of Education and Management Engineering* 7 (6): 40–49.

Künn-Nelen, Annemarie. 2016. "Does commuting affect health?" *Health economics* 25 (8): 984–1004.

Kuznetsova, Alexandra, Per B. Brockhoff, and Rune H. B. Christensen. 2017. "lmerTest Package: Tests in Linear Mixed Effects Models." *Journal of Statistical Software* 82 (13): 1–26.

Lei, Cheng, and Kin Fun Li. 2015. "Academic performance predictors." In *2015 IEEE 29th International Conference on Advanced Information Networking and Applications Workshops*, 577–581. IEEE.

Li, Fan, Alan M. Zaslavsky, and Mary Beth Landrum. 2013. "Propensity score weighting with multilevel data." *Statistics in medicine* 32 (19): 3373–3387.

Moro, Esteban, Dan Calacci, Xiaowen Dong, and Alex Pentland. 2021. "Mobility patterns are associated with experienced income segregation in large US cities." *Nature communications* 12 (1): 4633.

Nadaraya, Elizbar A. 1964. "On estimating regression." *Theory of Probability & Its Applications* 9 (1): 141–142.

Neyman, J. 1938. "Contribution to the Theory of Sampling Human Populations." *Journal of the American Statistical Association* 33 (201): 101–116.

Pebesma, Edzer. 2018. "Simple Features for R: Standardized Support for Spatial Vector Data." *The R Journal* 10 (1): 439–446. https://doi.org/10.32614/RJ-2018-009.

Pindek, Shani, Winny Shen, and Stephanie Andel. 2023. "Finally, some "me time": A new theoretical perspective on the benefits of commuting." *Organizational Psychology Review* 13 (1): 44–66.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rastrollo-Guerrero, Juan L., Juan A. Gómez-Pulido, and Arturo Durán-Domínguez. 2020. "Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review." *Applied Sciences* 10 (3).

Rodriguez-Hernandez, Carlos Felipe, Eduardo Cascallar, and Eva Kyndt. 2020. "Socioeconomic status and academic performance in higher education: A systematic review." *Educational Research Review* 29: 100305.

Rosenbaum, Paul, and Donald Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies For Causal Effects." *Biometrika* 70: 41–55.

Rubin, Donald B. 2005. "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions." *Journal of the American Statistical Association* 100 (469): 322–331.

Schuler, Megan S., Wanghuan Chu, and Donna Coffman. 2016. "Propensity score weighting

for a continuous exposure with multilevel data." *Health Services and Outcomes Research Methodology* 16: 271–292.

Serra, Márcia Filipa Alves. 2021. "The Impact of Commuting on Higher Education Students´ Academic Performance." PhD diss., Universidade NOVA de Lisboa (Portugal).

Shahiri, Amirah Mohamed, Wahidah Husain, and Nur'aini Abdul Rashid. 2015. "A review on predicting student's performance using data mining techniques." *Procedia Computer Science* 72: 414–422.

Silva, Ana Daniela, Jaisso Vautero, and Camilo Usssene. 2021. "The influence of family on academic performance of Mozambican university students." *International Journal of Educational Development* 87: 102476.

Sothan, Seng. 2019. "The determinants of academic performance: evidence from a Cambodian University." *Studies in Higher Education* 44 (11): 2096–2111.

Stopher, Peter, Camden FitzGerald, and Min Xu. 2007. "Assessing the accuracy of the Sydney Household Travel Survey with GPS." *Transportation* 34: 723–741.

Tigre, Robson, Breno Sampaio, and Tatiane Menezes. 2017. "The impact of commuting time on youth's school performance." *Journal of Regional Science* 57 (1): 28–47.

Torenbeek, Marjolein, Ellen Jansen, and Cor Suhre. 2013. "Predicting undergraduates' academic achievement: the role of the curriculum, time investment and self-regulated learning." *Studies in Higher Education* 38 (9): 1393–1406.

Vegetabile, Brian G., Beth Ann Griffin, Donna L. Coffman, Matthew Cefalu, Michael W. Robbins, and Daniel F. McCaffrey. 2021. "Nonparametric estimation of population average dose-response curves using entropy balancing weights for continuous exposures." *Health Services and Outcomes Research Methodology* 21: 69–110.

Vieira, Carlos, Isabel Vieira, and Luis Raposo. 2018. "Distance and academic performance in higher education." *Spatial Economic Analysis* 13 (1): 60–79.

Watson, Geoffrey S. 1964. "Smooth regression analysis." *Sankhyā: The Indian Journal of Statistics, Series A* 359–372.

Wilhoit, Elizabeth D. 2017. "'My drive is my sacred time': commuting as routine liminality." *Culture and Organization* 23 (4): 263–276.

Xia, Tianqi, Xuan Song, Haoran Zhang, Xiaoya Song, Hiroshi Kanasugi, and Ryosuke Shibasaki. 2019. "Measuring spatio-temporal accessibility to emergency medical services through big GPS data." *Health & Place* 56: 53–62.

Yousef, Darwish Abdulrahman. 2019. "Determinants of the academic performance of undergraduate students in statistics bachelor's degree program: A study in an Arabic setting." *Quality Assurance in Education* .

Zhu, Yeying, Donna L. Coffman, and Debashis Ghosh. 2015. "A boosting algorithm for estimating generalized propensity scores with continuous treatments." *Journal of causal inference* 3 (1): 25–40.

Zuluaga-Ortiz, Rohemi, Enrique DelaHoz-Dominguez, and Alicia Camelo-Guarín. 2022. "Academic efficiency of engineering university degrees and its driving factors. A PLS-DEA approach." *Journal of International Studies* 15 (2).

Appendix A. Trajectory estimation

The dataset provided by Cuebiq is extensive and has a wide spatiotemporal coverage. When limited to the metropolitan area of Milan and working days, it contains 26,397,613 GPS signals from 165,582 users' smartphones, describing their position in time both during spatial movements and stopovers. Among all this information, however, the current research is interested in obtaining measurements of time-to-university, i.e., just the ones related to journeys (and not stops) directed toward the university site. The wide original dataset is hence reduced and further elaborated to extract this valuable information and make it usable for accessibility map constructions.

The raw dataset is composed by GPS signals $P_i = (x_{1i}, x_{2i}, t_i)$, representing the i-th GPS sample measured at time t_i at UTM coordinates x_{1i} and x_{2i}. The observation P_i describes either the situation of a moving subject (e.g., the user is traveling from one place to another), the situation of a temporarily non-moving subject that will restart its journey (e.g., the user is waiting for the bus), or the situation of a permanently non-moving subject (e.g., the user is at home). The first two cases are associated with movements and their GPS signals are called moving points, while data associated with non-moving status are called stopping points.

As a first step, it is needed to elaborate and filter the original dataset of raw GPS signals $\{P_i\}_i$ to create a new version where stopping points are discarded and the remaining signals are grouped together when they describe the same spatial movement of the same user. The group of signals is the so-called journey, also known as trip or trajectory. A journey describes the spatial movements of a single user through a sequence of GPS data points, ordered with respect to the timestamp, which starts from a moving point and ends at a stopping point. Following the same notation as before, the k-th journey is characterized by the sequence of n_k points from the first one after a set of stopping points $P_{m_1^k}$ until the occurrence of a new stopping point $P_{m_{n_k}^k}$, and can be written as $M_k = \{P_{m_1^k}, \ldots, P_{m_{n_k}^k}\}$.

In order to assess the identification of trajectory, it is necessary to define standards for the detection of stopping points. These guidelines originate from the general and intuitive notion of stopping point, and are then formalized with the application of precise quantitative criteria. Consider two consecutive GPS signals of the same user $P_a = (x_{1a}, x_{2a}, t_a)$ and $P_b = (x_{1b}, x_{2b}, t_b)$, with $t_b > t_a$, and define the distance traveled $\Delta x = \sqrt{(x_{1b} - x_{1a})^2 + (x_{2b} - x_{2a})^2}$, the elapsed time $\Delta t = t_b - t_a$, and the average velocity $v = \Delta x / \Delta t$ between the two measurements. The selected identification criteria for stopping points set global constraints on distance, time, and velocity, applicable for all users, and additional device type-based constraints. Global criteria assess the measurements between consecutive instances and highlight the stability of locations over time through the introduction of thresholds on distance ($\Delta x = 0$ m), time ($\Delta t > 60$ min), and velocity ($v < 0.7$ m/s) to identify stopping points. Additional device type-based constraints are introduced to account for the data-collecting procedure, which differs according to the user's smartphone operating system: iOS-based devices sample new measurements after movements of about 50 meters from the previously recorded position, while Android devices after approximately 5 minutes. Hence additional constraints are added on time and velocity ($\Delta t < 15$ min and $v < 1$ m/s) in the Android case and on distance and velocity ($\Delta x < 100$ m and $v < 1$ m/s) in the iOS case.

Once the stopping points are identified and the trajectories are found, the full dataset of $\{M_k\}_k$ is reduced to consider just the journeys that end in the proximity

of the two university campuses: if the final stopping point of a journey is at most 250 meters away from the site, then the trajectory is kept; otherwise, it is discarded. The reduction is needed to get a GPS dataset representative of the mobility pattern to any point in Milan to the university site, which mimics the behavior of a student traveling for attending lectures. Further filtering is applied to ensure more robustness in the measurements, removing journeys with less than 6 observations, journeys with at least one signal of low accuracy ($acc \geq 1,500$ m), and self-loop journeys from and to the same university campus. Upon the conclusion of this phase, the dataset pertaining to the Leonardo campus consists of 7,670 GPS measurements of 881 trajectories from 448 unique Cuebiq users. Trajectories are composed of 8.6 signals on average and have a mean length of 7.8 kilometers, ranging from 300 meters to a maximum of 33 kilometers. More than 62% of Cuebiq users are associated with only one trajectory, almost 18% with two trajectories, and the few remaining can reach 10 trajectories. Similarly, in relation to the Bovisa campus, there are 4,111 GPS measurements associated with 498 trajectories from 295 distinct Cuebiq users. The dataset is almost half the size of the previous one. On average, with respect to the case of Leonardo campus, trajectories result longer in terms of distance traveled (8.7 kilometers on average) and shorter in terms of number of signals (7.92 on average) and user (1.6 trajectories per user on average). All this information is summarized in Table A1.

Finally, the commuting time to university of each GPS measurement is computed as the amount of time spent from its measurement to the end of the trajectory. Let us consider the k-th journey M_k. The commuting time to university y_{m^k} associated to the point $P_{m^k} = (x_{1m^k}, x_{2m^k}, t_{m^k})$ for any $m = m_1^k, \ldots, m_{n_k}^k$ is computed as $y_{m^k} = t_{m_{n_k}^k} - t_{m^k}$.

Table A1. Descriptive statistics of the GPS dataset after elaboration and filtering for both the campus of Leonardo and Bovisa

Descriptive statistics	Leonardo campus	Bovisa campus
n. signals	7,670	4,111
n. trajectories	881	498
n. signals per trajectory	[6; 85] mean=8.6	[6; 55] mean=7.92
distance per trajectory [km]	[0.3; 33.0] mean=7.8	[0.6; 36.5] mean=8.7
n. unique users	448	295
n. signals per user	[6; 183] mean=15.7	[6; 149] mean= 13.9
n. trajectories per user	[1; 10] mean=1.8	[1; 8] mean=1.6

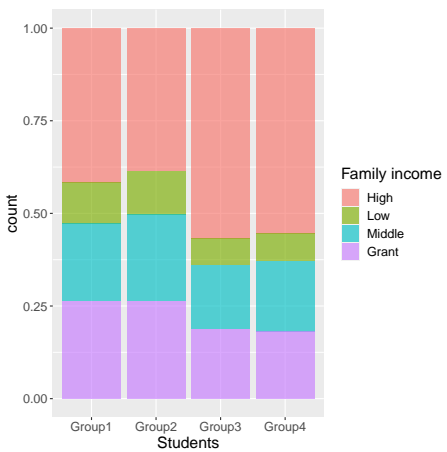## Appendix B. Qualitative analysis on the representativeness of the students' sample

The sample students selected for the analyses are a small portion of the entire population of first-year engineering students at Politecnico di Milano. Indeed, there are
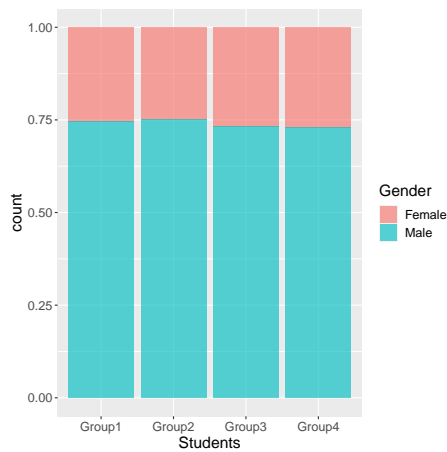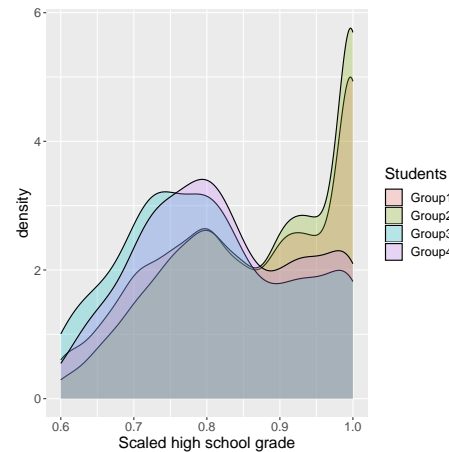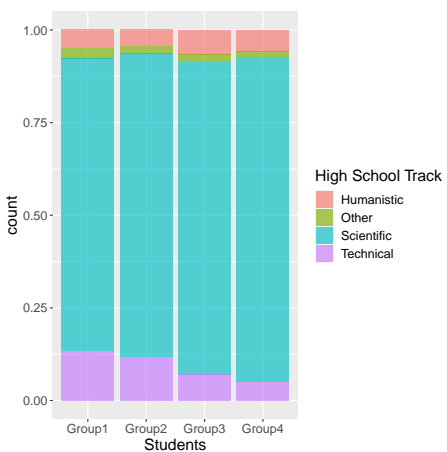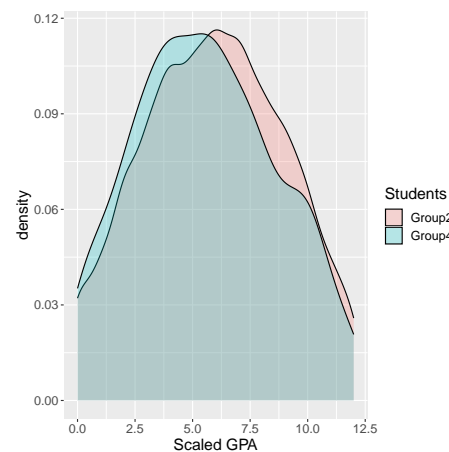
(a) Gender
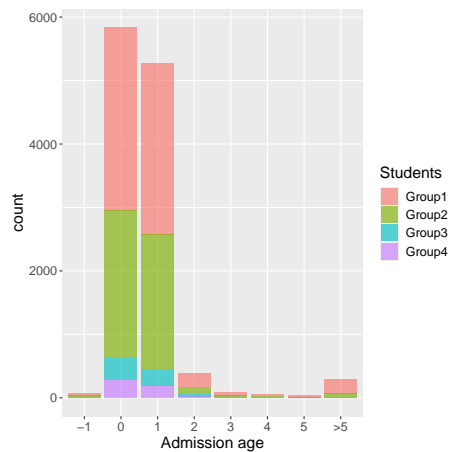
(b) Admission age

(c) Family income

(d) High school grade

(e) High school track

(f) GPA

Group 1: first-year Engineering Bachelor students in a.y. 2019/20; Group 2: first-year Engineering Bachelor students in a.y. 2019/20 who pass at least one exam; Group 3: first-year Engineering Bachelor students in a.y. 2019/20 resident in Milan; Group 4: first-year Engineering Bachelor students in a.y. 2019/20 resident in Milan who pass at least one exam.

Figure B1. Comparison of samples distributions

considered only students who passed at least one exam and whose residence location is in the Milan metropolitan area. This fact may limit the interpretation of the results to the only study case, thus preventing their generalization to the entire student body. Hence, this section aims to demonstrate the representativeness of the selected sample and to assess the degree to which the findings can be extended. Firstly, a comparison between the groups of students is carried out by looking at the summary statistics of the variables. Then, focusing on the restriction of students living in Milan, a regression model is used to investigate the relationship between the indicator of passing and not at least one exam and the commuting time.

More formally, four distinct groups of students are defined. Group 1 represents the entire population of first-year students enrolled in the bachelor engineering program at the Politecnico di Milano during the first semester of the academic year 2019/20, with a total of 6166 students. Group 2, comprising 4692 students, includes those students in Group 1 who have successfully completed at least one exam by the end of the first semester. The sample is then restricted to students residing in Milan. Group 3

Table B1. Comparison of samples descriptive statistics

Variable's name	Group1[a] N=6166	Group2[b] N=4692	Group3[c] N=669	Group4[d] N=507
Gender				
Male	4596 (74.5%)	3524 (75.1%)	490 (73.2%)	370 (73.0%)
Female	1570 (25.5%)	1168 (24.9%)	179 (26.8%)	137 (27.0%)
Admission age				
Range	[-1;37]	[-1;30]	[-1;14]	[-1;14]
Mean	1.022	0.66	0.67	0.57
SD	2.69	1.39	1.23	1.11
Family income				
High	2567 (41.6%)	1809 (38.6%)	380 (56.8%)	281 (55.4%)
Middle	1288 (20.9%)	1091 (23.3%)	115 (17.2%)	96 (18.9%)
Low	686 (11.1%)	554 (11.8%)	48 (7.2%)	38 (7.5%)
Grant	1625 (26.4%)	1238 (26.4%)	126 (18.8%)	92 (19.2%)
High school grade				
Range	[0.6;1]	[0.6;1]	[0.6;1]	[0.6;1]
Mean	0.85	0.87	0.81	0.82
SD	0.12	0.11	0.12	0.11
High school track				
Humanistic	300 (4.9%)	203 (4.3%)	44 (6.6%)	29 (5.7%)
Scientific	4864 (78.9%)	3848 (82.0%)	565 (84.5%)	444 (87.6%)
Technical	829 (13.5%)	552 (11.8%)	47 (7.0%)	26 (5.1%)
Other	173 (2.8%)	89 (1.9%)	13 (1.9%)	8 (1.6%)
Commuting time				
Range			[0;0.76]	[0;0.76]
Mean			0.38	0.38
SD			0.17	0.17
GPA				
Range		[0;12]		[0;12]
Mean		5.9		5.6
SD		3.02		3.04

[a]First-year Engineering Bachelor students in a.y. 2019/20; [b]First-year Engineering Bachelor students in a.y. 2019/20 who pass at least one exam; [c]First-year Engineering Bachelor students in a.y. 2019/20 resident in Milan; [d]First-year Engineering Bachelor students in a.y. 2019/20 resident in Milan who pass at least one exam.

consists of those 669 individuals of Group 1 who live in Milan (i.e., first-year bachelor students in engineering at Politecnico di Milano living in Milan). Parallelly, Group 4 encompasses 507 students of Group 2 who are residing in Milan (i.e., first-year bachelor students in engineering at Politecnico di Milano living in Milan and passed at least one exam by the end of the first semester), and composes the sample used in the analyses. Note that for students in Groups 2 and 4 it is possible to compute the GPA, while the remaining students are not associated with any mark.

Table B1 shows the main descriptive statistics. With the same objective, Figure B1 illustrates the distributions of the analyzed variables in the four samples. No major differences emerge with regard to gender and age of admission between different groups. However, disparities are noted in terms of household income, with apparently higher income among Milan residents. In addition, the latter appear to have taken more science and humanities and less technical courses during high school, and have obtained lower final grades in high school. Finally, in terms of GPA, although similar mean values occur in the two samples, the GPA distribution seems slightly shifted to smaller values for Group 4 than for Group 2.

Finally, the study employs a Generalized Linear Mixed-effect Model (GLMM) to examine the association between the variable indicating whether a student passes at least one exam at the end of the first semester or not and the commuting time, restricting the analysis to students who live in Milan. The goal of the analysis is to highlight the relevance of commuting time in distinguishing between students passing or not an exam, i.e., belonging to Group 3 or Group 4. For each student i, $i = 1, \ldots, n_l$ in the program l, $l = 1, \ldots, L$, $n = \sum_l n_l$, the model can be written as:

$$y_{il} \sim Be(p_{il})$$
$$logit(p_{il}) = \beta_0 + \sum_{j=1}^{J-1} \beta_j x_{jil} + \beta_J a_{il} + u_l \tag{B1}$$
$$u_l \sim N(0, \sigma_{u_2}^2)$$

where y_{il} is the output variable of student i (1 if they pass at least one exam), in program l; $\boldsymbol{\beta} = \{\beta_0, \ldots, \beta_J\}$ is the $(J + 1)$-dimensional vector of parameters; x_{jil} is the value of the j-th predictor at student's level; a_{il} represents the commuting time; u_l is the random effect of the program l.

The model successfully classifies 508 out of 669 students using LOO-CV, achieving an accuracy of 76%. Figure B2a presents the Odds Ratio values, indicating that both high school grade and the specific high school attended are statistically significant variables. Although the differences in other variables, it is important to note that the variable of main interest, i.e., commuting time, is not statistically significant.

In light of the exploratory analysis involving descriptive statistics, comparisons of sample distributions, and model evaluations, it can be confidently concluded that the sample is appropriate to describe students living in the metropolitan area of Milan.
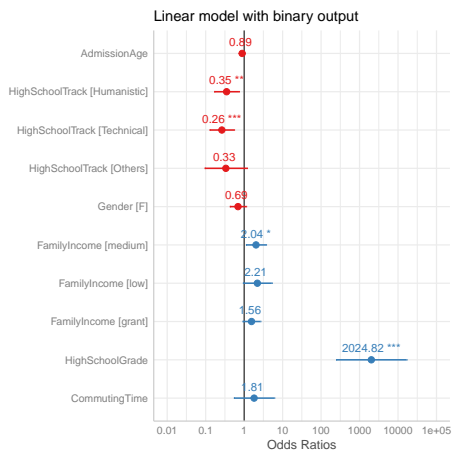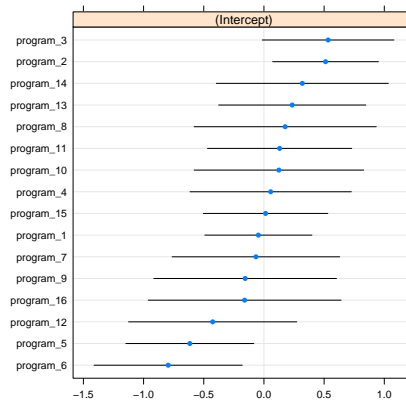
(a) Odds Ratio   (b) Random effects

Figure B2. Estimated coefficients of the model with binary output (Mod. B1).

MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

91/2023   Burzacchi, A.; Rossi, L.; Agasisti, T.; Paganoni, A. M.; Vantini, S.
Commuting time as a determinant of higher education students' performance: the case of Politecnico di Milano

90/2023   Gregorio, C.; Baj, G.; Barbati, G.; Ieva, F.
Dynamic treatment effect phenotyping through functional survival analysis

88/2023   Masci, C.; Spreafico, M.; Ieva, F.
Joint modelling of recurrent and terminal events with discretely-distributed non-parametric frailty: application on re-hospitalizations and death in heart failure patients

89/2023   Savaré, L.; Ieva, F.; Corrao, G.; Lora, A.
Capturing the variety of clinical pathways in patients with schizophrenic disorders through state sequences analysis

86/2023   Ferraccioli, F.; Sangalli, L.M.; Finos, L.
Nonparametric tests for semiparametric regression models

85/2023   Arnone, E.; De Falco, C.; Formaggia, L.; Meretti, G.; Sangalli, L.M.
Computationally efficient techniques for Spatial Regression with Differential Regularization

83/2023   Cavinato, L.; Massi, M.C.; Sollini, M.; Kirienko , M.; Ieva, F.
Dual adversarial deconfounding autoencoder for joint batch-effects removal from multi-center and multi-scanner radiomics data

82/2023   Pozzi, G.; Ciarletta, P.
Geometric control by active mechanics of epithelial gap closure

81/2023   Buchwald, S.; Ciaramella, G; Salomon, J.; Sugny, D.
A SPIRED code for the reconstruction of spin distribution

79/2023   Agosti, A.; Bardin, R.; Ciarletta, P.; Grasselli, M.
A diffuse interface model of tumour evolution under a finite elastic confinement