Mixed Data Sampling in Learning Analytics

Julian Langenhagen
Goethe University Frankfurt, Germany
langenhagen@econ.uni-frankfurt.de

ABSTRACT

Technical progress facilitates collecting large amounts and new kinds of data in a wide range of areas. That enables versatile new possibilities in empirical research, especially with high-frequency data. However, researchers are confronted with the problem that not all available data have the same (high) frequency. For many common methods, it is necessary to adjust the high-frequency to the low-frequency data, resulting in a significant loss of information. In accounting research, this kind of problem exists due to the low-frequency reporting data of companies on the one hand and the high-frequency financial market data on the other. A promising solution to this problem is the innovative approach of mixed data sampling (MIDAS). Since the coexistence of low-frequency data (e.g., exam grades) and high-frequency data (e.g., learning management system usage data) is also prevalent in educational settings, this paper will discuss the first application of MIDAS in the field of learning analytics.

Keywords

time series, mixed data sampling, regression, prediction models

1. INTRODUCTION

Educational Data Mining (EDM) is a comparatively young field of research. Even though certain research methods within this area are already established, the field regularly benefits from valuable contributions from interdisciplinary research approaches [e.g., 14]. The methods used in Educational Data Mining can be divided into four different areas: prediction models, structure discovery, relationship mining, and discovery with models [2]. The focus of this paper lies on prediction models, especially on those with time-series data. Methods in this area include classifications, regressions, and latent knowledge estimation [2]. In this area, EDM researchers are confronted with a specific problem. The variety of available data sources is growing due to the use of complex learning management systems (LMS), game-based learning applications, or other digital educational tools. These sources often contain high-frequency data and therefore offer a lot of potential information. However, in the most commonly used methods in prediction models, it is usually the case that data from different samples have to be brought to the same frequency to be analyzed. This can lead to a significant loss of information. For example, data from an LMS can be gathered for every possible usage second to be used as an explanatory variable. However, there is usually a low-frequency variable on the other side of the equation, such as the exam grade or score. Accounting researchers face a similar problem. Companies usually only publish reports at a pre-defined low frequency, as financial statements or other reports are often made available only annually or quarterly. An independent variable available in corresponding research settings is, for example, the share price, which, like the data from the LMS, can be collected at a very high frequency. However, the information contained here cannot be fully included in the analysis if the variables on both sides of the equation have to be adjusted to the lowest frequency available. A solution to this problem in accounting is the method of mixed data sampling (MIDAS). As the underlying problem is comparable to typical educational research settings, this method will be examined in more detail in the following section, and then a possible application in Educational Data Mining will be discussed using the example of a concrete implementation in the context of a learning app in higher education.

2. MIDAS IN ACCOUNTING RESEARCH

The basic MIDAS model builds on a regression equation where the dependent variable is measured in a lower frequency than one or more of the independent variables [5]. The problem of the different frequencies on both sides is solved with two separate components. In the first component, each time disaggregated observation of the higher frequency variable is included separately as an independent variable. In other words, if the higher frequency data is observable N times within a period, then N separate independent variables are included in the regression. This allows the independent variable's effect on the dependent variable to evolve over the course of the examined period, even though the dependent variable was only measured once. The second component of MIDAS is the requirement of each of the N regression coefficients to follow a specific function of time that is shaped by few estimated parameters. For example, if the temporal distribution is assumed to be linear, this condition only requires two parameters, namely an intercept and a slope. This condition is the key feature of MIDAS to establish a balance between model flexibility and parsimony to be able to reasonably interpret the results. This basic model can be enhanced in many different directions, e.g., to whether certain events within the observed period have a different relationship to non-event data [5] or for the evaluation of unequally spaced temporal data [11]. MIDAS has so far been used mainly in accounting and macroeconomics research [e.g., 4, 5, 8]. The method is particularly well suited to accounting research as companies are legally obliged to publish certain economic data such as revenues and costs on a regular low-frequency cycle (e.g., quarterly or annually). Share prices, on the other hand, can generally be retrieved every second and are therefore high-frequency. The job of professional analysts is to use this information, among other things, to predict the companies' disclosures. For forecasts based on regressions, it is usually necessary to adjust the frequency of the different data samples to the lowest frequency available. For a share price, this could result in the quarterly mean, for example. Therefore, the information within the high-frequency stock market data lost in this process represents a major challenge for analysts to optimize their forecasts. A previous study has shown that MIDAS can significantly help analysts with this challenge and improve the forecasts accordingly [6]. Building on these findings, the next section will discuss whether MIDAS can also help lecturers and researchers in education with specific predictions.

3. MIDAS IN LEARNING ANALYTICS

Prediction models belong to the most used methods in Educational Data Mining [1, 7]. Usually, the corresponding analyses are carried out with classifications or regressions [3]. The prediction of exam grades or scores is one of the most frequently investigated research questions within prediction models [13]. The dependent variables are usually the exam points (regression), the exam grade (classification or regression), or the fact of whether a student has passed or not (classification). In many cases, usage data from learning management systems are used as independent variables. This data is usually high-frequency and must be adjusted to the low-frequency dependent variables for the methods mentioned above. For example, an LMS may record all clicks within the system with precise temporal data. Still, in the above analyses, this information needs to be restricted, for example, to the total number of clicks over the entire period of use [9]. These limitations could be overcome with more sophisticated methods. For instance, it was shown that GARCH, a method from finance research, outperformed the other common methods in multi-modal learning analytics [14]. As of this writing, there is no publication in the field of Educational Data Mining or Learning Analytics that has used the MIDAS approach. This research gap should be filled with the present project. In a subsequent step, MIDAS could even be linked to GARCH to further enrich the research setting [10]. Thus, the research question of this project is whether MIDAS is suitable for predicting exam results in an educational context and how the results compare to those of already known methods in Educational Data Mining. Previous studies have shown that it is important to look not only at aggregate usage data for a given time period but also at the distribution and sequence of the corresponding data points [e.g., 9, 12]. MIDAS could make a valuable contribution to the range of methods already available, as it takes into account the high frequency of independent variables while still providing well-interpretable and thus actionable results for instructors. These results could, for example, be used to build an early warning system for students at risk of academic failure. The good interpretability of MIDAS results could make it easier for instructors to take appropriate measures compared to when using complex machine learning algorithms, whose results might be much more challenging to interpret. Such an early warning system is especially beneficial in lectures where there is little performance feedback between students and teachers in general (e.g., because there is only one final exam at the end of the semester) or due to special conditions (e.g., COVID-19). In such cases, all actors involved see the result of learning and teaching behavior only at the end of the semester through the exam result. Since it is already too late for countermeasures at this point, an early warning system with clear recommendations for action would be beneficial in such a context. Therefore, MIDAS is a promising addition to the current variety of methods and should be considered in future studies.

4. NEXT STEPS

A first application of the basic MIDAS model will be carried out in the following setting as soon as the project's data collection is completed. We developed a mobile learning app for an undergraduate accounting course at a large public university in Europe. The course is compulsory and ought to be taken in the third semester of the bachelor's program. The course is taken by approximately 600 students per semester and consists of a weekly lecture, a biweekly exercise, and biweekly tutorials (five meetings in small groups). The content of this course includes the basics of cost accounting as well as a summary of their significance and classification in the management accounting context. The primary learning material consists of a slide deck, a collection of exercises (with solutions), and a trial exam (all available as PDF files). In the evaluations of earlier semesters, students often complained that there were no contemporary possibilities to learn the subject matter. Therefore, we decided to develop an additional learning tool in the form of a smartphone app, which was launched in the summer semester of 2019. The use of the app is voluntary, and no extra credits or advantages for the final exam can be earned by collecting points in the app. The tool is available via a web version and as an app in the Google Play Store and the Apple App Store. The app's core element is a database with over 550 questions that covers all nine chapters of the course. In addition to the question types single and multiple-choice, there are also sorting and cloze text tasks. The app can be used in three modes: The chapter mode can be used to answer specific questions about a single chapter. As soon as a student has mastered the problems of one chapter, the next chapter is unlocked. In random mode, questions are randomly selected from the chapters that have already been unlocked in chapter mode. In the third mode, the so-called Weekly Challenge, users can compare themselves with other students. Once a week, they have the opportunity to answer 25 questions randomly selected from the chapters already covered in the lecture. The results are subsequently displayed in a weekly and a semester ranking. For good performances in the Weekly Challenge and other learning achievements, students can earn so-called badges, which are then displayed in their account under their self-chosen username. By answering questions (regardless of the mode), students also earn learning points and thus increase their learning level. The progress display of the individual chapters shows students how well they currently master a particular topic. The app has been specifically designed to complement the existing course and is not intended to replace other learning materials such as the slides or the collection of exercises. The app contains an individual explanation for each question which is displayed if a wrong answer is given. Thus, students can work their way through the catalog of questions independently of time and place and eliminate any gaps in their understanding without having to rely on the presence of the lecturers. This is an essential value-added for the students, especially in such a large course with approximately 600 students per semester. The collected app data consists of details about the usage behavior of each student (e.g., time of use, performance (history) regarding every question, and earned badges). At this stage, we already have four semesters of app usage, and the data set is growing as the research project is still ongoing. This is especially promising as the situation regarding COVID-19 lead to an exogenous shock. While in the years 2018 and 2019, the course was held face-to-face, in the summer semester 2020, it was converted into a purely online lecture. Apart from the launch of the app and the switch to an online lecture, there were no teaching design changes over the course of the semesters. The lecturer and the learning materials remained constant, as well as the design and the grading of the final exam. This unique setting could provide valuable insights into the impact of COVID-19 on higher education. The starting point of the corresponding analysis would be a basic linear regression with the exam score as dependent and usage data from the app as independent variable. The exam score is measured once, while the usage data from the app could be evaluated by every second of the semester. In this setting, we face the challenge of unequal frequencies on both sides of the equation that was described before. If we would only take the sum of total questions answered by a student as the independent variable, we would miss a lot of information. The type and especially the time of usage can be decisive for the effect on the exam score. We would miss all this information with reducing the app usage on measures like total questions answered. Therefore, the MIDAS approach offers a promising possibility to extract more insights from the data set. A comparative analysis with other already known methods in Educational Data Mining or Learning Analytics, which take into account the high frequency of data, could highlight the additional benefits of MIDAS for this research area.

5. CONCLUSION AND FUTURE WORK

In this paper, it was shown that the innovative approach MIDAS could be a promising extension of the variety of methods in Educational Data Mining and Learning Analytics. Further insights will be gained by testing the approach with the usage data from our gamified learning app. Based on the findings, it will be further discussed whether the basic model of MIDAS should be extended for the use in an educational setting or whether other novel methods should be applied in this setting. Besides, it could be promising to apply the MIDAS approach to already published analyses in order to test the corresponding added value. If the results are similarly insightful as those in accounting research, MIDAS could find numerous use cases in Educational Data Mining and Learning Analytics.

6. REFERENCES

[1] Aldowah, H. et al. 2019. Educational data mining and learning analytics for 21st century higher education: A review and synthesis. *Telematics and Informatics*. 37, (2019), 13–49.

[2] Baker, R.S. and Inventado, P.S. 2014. Educational data mining and learning analytics. *Learning analytics*. Springer. 61–75.

[3] Bakhshinategh, B. et al. 2018. Educational data mining applications and tasks: A survey of the last 10 years. *Education and Information Technologies*. 23, 1 (2018), 537–553.

[4] Ball, R.T. et al. 2019. Tilting the evidence: the role of firm-level earnings attributes in the relation between aggregated earnings and gross domestic product. *Review of Accounting Studies*. 24, 2 (2019), 570–592.

[5] Ball, R.T. and Gallo, L.A. 2018. A mixed data sampling approach to accounting research. *Available at SSRN 3250445*. (2018).

[6] Ball, R.T. and Ghysels, E. 2018. Automated earnings forecasts: beat analysts or combine and conquer? *Management Science*. 64, 10 (2018), 4936–4952.

[7] Chen, G. et al. 2020. Let's shine together! a comparative study between learning analytics and educational data mining. *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge* (2020), 544–553.

[8] Clements, M.P. and Galvão, A.B. 2009. Forecasting US output growth using leading indicators: An appraisal using MIDAS models. *Journal of Applied Econometrics*. 24, 7 (2009), 1187–1206.

[9] Conijn, R. et al. 2016. Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS. *IEEE Transactions on Learning Technologies*. 10, 1 (2016), 17–29.

[10] Engle, R.F. et al. 2013. Stock market volatility and macroeconomic fundamentals. *Review of Economics and Statistics*. 95, 3 (2013), 776–797.

[11] Ghysels, E. et al. 2007. MIDAS regressions: Further results and new directions. *Econometric reviews*. 26, 1 (2007), 53–90.

[12] Malekian, D. et al. 2020. Prediction of students' assessment readiness in online learning environments: the sequence matters. *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge* (2020), 382–391.

[13] Romero, C. and Ventura, S. 2010. Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*. 40, 6 (2010), 601–618.

[14] Sharma, K. et al. 2019. Modelling Learners' Behaviour: A Novel Approach Using GARCH with Multimodal Data. *European Conference on Technology Enhanced Learning* (2019), 450–465.