Online data repositories as educational resources? A learning environment covering formal and informal inferential statistics ideas in scientific inquiry

Thomas Schubatzky and Claudia Haagen-Schützenhöfer

University of Graz, Institute of Physics, Department of Physics Education, Austria

E-mail: thomas.schubatzky@uni-graz.at

Received 20 December 2018, revised 22 March 2019
Accepted for publication 8 May 2019
Published 18 June 2019

Abstract

Statistical ideas play a vital role in scientific investigations. For students enrolling in physics-related courses at university, the need to interpret data is set from the start. Analysing graphical representations of data is seen as one way to acquaint students with statistical thinking without relying on pre-knowledge about formal statistics for applying more sophisticated methods like multiple regression. We designed a learning environment, which supports students in understanding the exploratory analysis of multivariate datasets as well as the concept of multiple regression. For phase one of the learning path, we work with exploratory data analysis using the software TinkerPlots, which has several major advantages in contrast to conventional software. Only in phase two formal inferential statistics is applied. We have chosen a context-oriented approach for this learning environment, using the particulate matter concentration in an Austrian city as topic. By providing data from online data repositories in a simplified way, students get the opportunity to work with real data. The amount of this data exceeds the number of measurements collected in typical training labs in an authentic and feasible way. In this article, the design of the intervention and a range of results originating from the triggered learning paths will be presented and discussed. To sum it up, we illustrate advantages and opportunities of the use of innovative software, online data repositories and informal statistics for a first introduction of methods of formal statistics like multiple regression.
Supplementary material for this article is available online

Keywords: informal inferential statistics, particulate matter, TinkerPlots, online data repositories, scientific inquiry, multiple regression analysis

(Some figures may appear in colour only in the online journal)

1. Introduction

Data plays a crucial role in today’s society. Not only because drawing inferences from data is part of everyday life, but also because critically reviewing results of statistical inferences from (science-) research studies is an important goal for students enrolling at a university.

In a science and technology-related studies or when studying physics as a minor, the need to interpret data and to draw conclusions from data is set from the get-go, but methods of formal inferential statistics are often either taught at a later stage in the curriculum or completely neglected. To prepare students for statistics courses in the first semesters and at the same time foster students’ ability to make inferential statements about data, learning environments using graphical representations can be one way to achieve this goal.

In today’s world, many research groups and governmental agencies make their research or environmental data accessible via online data repositories or by providing data at official homepages. For example, this can be data from an experiment conducted during research, meteorological data like temperature, humidity or wind velocity but also environmental data like air pollution (particulate matter, NOX, …) or light pollution. By providing university students with data from these repositories in a simplified way, students get the opportunity to work with real data that exceeds the amount of a few measurements—e.g. collected in lab courses—in an authentic and feasible way.

Overall, an example of how these two ideas—supporting the ability of informal inferential reasoning about data and using online data repositories—can be connected will be presented in this article. In the next section, the overall design of our learning environment covering those ideas is described. Subsequently, the underlying theoretical framework and ideas will be the central theme.

2. The learning environment

Our learning environment is meant to support freshman students’ statistical thinking and skills, which are relevant for scientific correlational investigations. For freshmen of physics-related studies there are usually courses on the introduction of physical measurement and methods of analysis in preparation for different lab courses, where our learning environment can be implemented as an introduction to multiple regression. On the other hand, we used this intervention in an introductory course to digital media and technology for future physics teachers. In this course they are supposed to learn how to plan and carry out scientific inquiries using modern technology themselves and how this can also be transferred to school settings. For this type of course we see even an additional benefit: the students are supported to reach learning goals in the field of analysing data in a physics context. In addition, at least the informal inferential statistic part of our intervention acquaints them with a methodology and tools which are also applicable in an educational school context e.g. for inquiry based learning. However, in this article the focus is on both: exploratory analysis of data but also on introduction to multiple regression analysis.
Hence, the overall goals of this learning environment are to enhance students’ informal inferential reasoning and to introduce students to the concept of multiple regression. However, students should not only learn about the mathematical underpinning, but also understand mind-sets that are involved when conducting correlational investigations in science.

For this reason, we chose a context-oriented approach, using a real-world context that is relevant to students. In our case, students need to solve a real-world problem in a city of Austria. Thereby, students’ main task is to figure out which factors influence particulate matter concentration using data from meteorological stations. In the first part, students are assigned a task covering this problem: they are experts in the ‘department for air monitoring’ of the municipal government and their goal is to first identify factors that influence particulate matter concentration in Graz (Part I to III in figure 1). Afterwards, they develop a model that can predict particulate matter concentration in Graz (Part IV to VI in figure 1), so the local government can provide solid information for the citizens. Their assignment consists of two major parts: as the result of the first stage of the learning environment, they have to present findings from an exploratory data analysis where they describe which factors influencing PM10 concentration they have identified. Then, an input phase follows which introduces the concept of multiple regression. Based upon their findings of the exploratory data analysis, the students need to construct a model using multiple regression, which is able to best predict particulate matter concentration in Graz in the next step. These models will be presented and discussed in the final session of our intervention (Part VI, see figure 1).

In total, this intervention lasts for about 10 hours when students have to perform all analysis in the course. If students perform parts of the analysis at home, it would be shorter. As far as pre-knowledge is concerned, students should have basic knowledge of descriptive statistics (mean, median, box-plots, …) and in the best case some knowledge about linear regression.

While students will definitely learn about particulate matter while performing their analysis, the main goals of this learning environment are: students should learn how to perform exploratory data analysis using a multivariate dataset and how to use informal inferential reasoning. Furthermore, they should learn how to apply multiple regression to a real-world problem. In the subsequent sections, we first describe the underpinning theoretical frameworks guiding the design of this learning environment. Afterwards, possible results from the exploratory data analysis phase and resulting multiple regression models are presented.
3. Statistical ideas in scientific inquiry

Statistical ideas are not only part of our everyday life, but also an integral part of every scientific inquiry. Especially ideas of inferential statistics, where statements about a population are inferred from a relatively small sample, are part of each scientific inquiry. For example, imagine a simple experiment: dropping two balls, where one ball has double the mass of the second ball. The two balls are dropped from the same initial position and the time the ball takes to touch the ground will be measured. The experiment is repeated for example ten times. When conducting this experiment, one will probably realize that the time it takes for the balls to reach the ground will be about the same for both balls but also for each measurement. When conducting this experiment, the conclusion will probably be that two objects, which are only separable by their mass, always take the same time to reach the ground. Therefore, in this example, the conclusion is not built upon methods of formal inferential statistics, but on key ideas of statistical thinking. However, the use of data as evidence is still involved when formulating conclusions. The Intergovernmental Panel on Climate Change (IPCC) for example even defined a ‘Guidance Note on Uncertainty’, where they describe how to address uncertainty in key findings (Mastrandrea 2010).

When looking at students’ ideas about scientific inquiry we frequently find a limited perspective reduced to a single method, ‘the scientific method’. Lederman et al (2013) argue that this may be due to the overemphasis on the classical experimental design in science instruction. The problem is that experimental research is not representative for all scientific investigations. Amongst experimental studies, other types of research, for example descriptive or correlational studies, are also an important part of scientific inquiry (Lederman et al 2013). In experimental investigations, planned interventions with the manipulation of variables are conducted to derive causal relationships. In correlational studies however, the goal is to describe relationships between variables, which may have been identified via anterior descriptive research. The preference of correlational over experimental design can stem from various reasons: Maybe the line of research is not at a stage yet where experimental design makes sense or it is simply not possible to conduct experimental research due to external factors (for example in astrophysics or climate physics).

Experimental inquiry has also been the focus of educational research in the past years (Furtak et al 2012), but there is still lack of learning environments that target the execution of correlational investigations. Here ideas of inferential statistics play a big role due to multiple confoundations by external factors.

As theoretical basis for the design of the learning environment, the PPDAC-cycle developed by Wild and Pfannkuch (1999) was selected due to the nature of correlational investigations. This Problem–Plan–Data–Analysis–Conclusion inquiry cycle stems from statistics education and covers all the steps involved when conducting inquiry. Still, we want to emphasize that this list is merely a prototypical set of steps involved in scientific investigations with students.

The investigative cycle displayed in figure 2: PPDAC-cycle adapted after Wild and Pfannkuch (1999) shows ways of thinking and activities during a PPDAC-cycle, which holds remarkable similarities to inquiry-cycles used in science education. Typically, an educational inquiry process starts with a given problem, which students are supposed to investigate with ‘scientific methods’. The idea of the Problem stage of the PPDAC-cycle is to turn vague ideas of the problem into specific questions, which can then be answered using data (Wild et al 2018).

The Data step is about obtaining the data, storing and cleaning it (Wild et al 2018). The focus of this learning environment is not on this step, therefore we use already obtained data
from an online data repository of a local government. Subsequently, the **Analysis** and the **Conclusion** step are about making sense of the data, abstracting and communicating what has been deduced from the data. Those methods can range from graphical representations or calculation of statistical key figures to the testing of (statistical) hypothesis or other methods. Wild and Pfannkuch (1999) also emphasize the role of contextual knowledge when formulating conclusions about an investigation. The interplay of knowledge about data and knowledge about context is also a key element of experimental investigations, which are always guided by contextual knowledge. Conclusions drawn from empirical data, for example from an experiment, are always informed by previous investigations and accepted scientific knowledge (Lederman et al 2014).

While in physics both methods—experimental and correlational—are used on a regular basis, physics teaching at school and partly also at university level focus generally on experimental investigations and highly controlled science experiments. The same is also true for the focus of physics education research; the way students learn how to conduct correlational investigations has not been the focus of research.

In our project, we build on the fact that statistical ideas, especially ideas of inferential statistics are involved in every scientific inquiry in both correlational and experimental investigations. Hence, approaches to support statistical thinking are necessary. This is what our learning environment is supposed to do. For the exploratory data analysis phase of our learning environment (Part I to III), we use the software TinkerPlots. In the next section, we describe how innovative technologies and software like TinkerPlots can enhance traditional classroom teaching of statistics.

### 4. Innovative technology and benefits to learning and teaching statistics for physics

In recent years, there has been a rise in the development of software, applets, shiny apps and tools, which are supposed to aid students’ understanding of statistical ideas. Typically, those
tools are of interest because of their interactivity. Particularly the software packages FATHOM (Finzer 2007) and TinkerPlots (Konold and Miller 2005, Konold 2007) have been used by various educators in order to improve their teaching of statistical concepts. Whereas FATHOM covers more advanced methods of statistics, TinkerPlots is outstanding regarding interactivity, speed of analysis and possibilities to generate and manipulate graphical representations. For example, Frischemeier and Biehler (2013) designed a statistics course for preservice mathematics teachers based on the use of TinkerPlots. TinkerPlots, which is used in parts of our learning environment, is a dynamic graphing software package created for students from a constructivist perspective. When using real data e.g. from online data repositories, they can be directly imported into TinkerPlots via spreadsheets. In contrast to other software (for example Excel), data are first represented in a random two-dimensional way as shown in figure 3. Students can then choose which variable (called ‘attributes’ in TinkerPlots) they wish to drag and drop to the plot. Dragging the icons in the plot left or right (up or down) creates more or fewer bins representing the data. For numerical attributes (variables), a continuous scale can be chosen. Additionally, there are tools available to represent data in plots, for example reference lines, hat-plots, box-plots and dividers but also group size and percent. In addition, values like means and medians can be displayed.

There are several aspects that make a software like TinkerPlots superior to conventional software like e.g. Microsoft Excel. Watson and Donne (2009) for example highlight the speed of analysis while using this software. Chance et al (2007) also commented that students are able to dig deeply into large datasets when using TinkerPlots. The extra time required for such explorations comes from the elimination of calculations or data organization in order to produce graphical representations. Overall, we are also using the software TinkerPlots in order to eliminate routine tasks, which are typically involved in exploratory data analysis phases, allowing more time for higher order thinking and learning.

Figure 3. Graphical representation of data in TinkerPlots.
5. Particulate matter concentration in Graz—description of the data sample

The generic term of particulate matter (PM) refers to the majority of airborne particles. The size of particles categorized as particulate matter usually ranges from 0.001 μm to 10 μm. A further differentiation is often made to group the particles according to their size. The most known and common categories of particulate matter are PM10 and PM 2.5 (Esworthy 2013). PM10 is the proportion of all airborne particles that can enter the human respiratory tract through the larynx during respiration. This determined subset of the total particles contains by definition 50% of the particles with an aerodynamic diameter that is less than 10 μm (Esworthy 2013, WHO 2013).

In addition to the composition and sources of particulate matter, effects of PM10 on the human health are of social interest. The effects of particulate matter and air pollutants in general are usually topic of epidemiological studies. According to studies published (e.g. Kim et al 2015), negative health effects of particulate matter on humans are no longer disputable. For example, the city of Graz (Austria) is considered the ‘particulate matter stronghold’ of Austria. The reduction in life expectancy is approximately 17 months (Schneider et al 2005). However, according to a WHO-database, it can be positively noted that the PM pollution in Austria (and this trend holds in general for European countries) has been steadily declining since 2003, totalling around 35% in the period from 2003 to 2018. However, due to the absence of a threshold under which negative effects of particulate matter on human health disappear, this topic is relevant not only in (particularly) vulnerable areas.

To fully understand the investigations displayed in the subsequent sections, there is a need to provide additional contextual knowledge about the situation in Graz. The reason why Graz is considered the ‘particulate matter stronghold of Austria’ stems from multiple sources. One source of high PM10 concentration in Graz is unfavourable weather conditions; namely, low wind speed, rare days with precipitation and many days with temperature inversion (Hörmann et al 2005).

Temperature inversion is the circumstance that air temperature increases with an increase in altitude. This means, that during inversion, there is a layer of warmer air above cooler air. For Graz, this means that if the temperature in Graz (350 m above sea level) is lower than the temperature on a mountain in the surroundings of Graz (Schöckl, 1445 m), one can assume that there is temperature inversion in Graz. So, the temperature gradient between the meteorological station on Schöckl ($T_{\text{high}}$) and the two meteorological stations in Graz ($T_{\text{low}}$) is used as an indicator for temperature inversion in Graz. Wind and precipitation (rain or snow) are known to favour low concentrations of PM10.

Besides meteorological influences, urban traffic also influences PM10 concentrations. According to studies (Hörmann et al 2005), traffic is responsible for 50%–70% of the anthropomorphic PM10 pollution in Graz. Therefore, measurements of two different meteorological stations in Graz are used in our learning environment: one is situated in a sector of Graz with high traffic load, the second in a rather traffic-calmed sector of Graz.

Overall, our multivariate dataset includes variables, which influence particulate matter concentration in Graz, but also variables which do not. Table 1 shows all the variables included in the dataset.

The data sample was sourced from an online data repository provided by the state government of Styria (luis.steiermark.at). The PM10 concentrations and the relating covariates stem from two different meteorological stations in Graz; one in the south where there is a lot of traffic and one in the north of Graz where there is comparably less traffic (see figure 4). The data was collected during the last three winter seasons (October to March) from 2014/2015 until 2017/18.
Figure 4. Location of the meteorological stations in Graz (Google Maps, 2018, maps.google.com).

Table 1. Variables included in the data sample.

| Variable                                                        | Type of variable                        |
|-----------------------------------------------------------------|------------------------------------------|
| Daily precipitation                                             | dichotomous (yes/no)                     |
| Daily average wind speed                                        | numeric (\(\text{m} \text{s}^{-1}\))     |
| Daily average temperature difference between Schöckl and the station in Graz | numeric (\(^\circ\text{C}\))            |
| Daily average PM10 concentration of the day before              | numeric (\(\mu\text{g m}^{-3}\))        |
| Type of day (Workday and Sun-/Holiday)                         | dichotomous (workday/sunday)             |
| Whether the day was in February or March                        | dichotomous (yes/no)                     |
| Daily average air humidity                                      | numeric (\(\%\))                        |
| Daily average temperature at the stations in Graz               | numeric (\(^\circ\text{C}\))            |
| Type of station                                                 | dichotomous (low traffic/high traffic)    |
6. Exploratory analysis of the factors influencing PM10 concentration in Graz using methods of informal inferential statistics

In this section, various factors influencing PM10 concentration are analysed by means of graphical representations. However, these graphs are just a selection of possible relationships students can identify while conducting their exploratory data analysis. With TinkerPlots there is no need to prepare the data or conduct further calculations in order to generate meaningful graphs, so students are able to dive deeply into the dataset in a fairly limited amount of time.

6.1. Analysing the influence of continuous variables

The relationship between inversion and PM10 concentration can be analysed in various ways. As a first task, bins with different values for the temperature gradient can be built. For example, four bins with a range of 6 °C per bin. In TinkerPlots, this can be done by selecting the relevant column in the data table (in this case the temperature gradient) and dragging one data point slightly to the right. Following this procedure, the desired number of bins can be displayed. After that, box-plots can be created for every bin with the box-plot function of TinkerPlots. Doing so, the quantiles and medians of the different bins can be compared. As figure 5 shows, we can observe a clear positive trend based on the temperature gradient by comparing the box-plots of the data used.

When dragging a data point even further to the right, the x-axis can also be transformed into a continuous axis. In doing so, a simple scatterplot can be generated as well.

6.2. Analysing mediation effects

Another strategy in analysing data sets is to check whether one factor influences the effect of another factor on the concentration of PM10. For this kind of analysis, TinkerPlots provides the option to colour-code the data points in a plot. The colour intensity of a data point represents different values.

Figure 5. Analysis of the relationship between average temperature gradient and PM10 concentration with box-plots.
Figure 6 shows a scatterplot where the data points are colour-coded concerning the average wind speed (ranging from deep purple (= high average speed) to white (= low average speed)).

The graph in figure 6 hints at a relationship between the average wind speed and both variables—the temperature gradient and the PM10 concentration.

Without inferential statistics, this result cannot be interpreted in a meaningful way without deeper contextual knowledge of the topic particular matter. As described above, high PM10 concentrations (in Graz) are usually linked to inversion. High wind speed is considered one of the main factors preventing inversion. So contextual knowledge about particulate matter and inversion are essential for a meaningful interpretation of the graph in figure 6: High average wind speed prevents inversion and is thus characterised by a low temperature gradient. This relationship can also be visualized with a scatterplot. In terms of statistics, the temperature gradient acts as mediator for the average wind speed when predicting PM10 as shown above.

This task shows that by using TinkerPlots, even more advanced statistical ideas like mediators can be discussed in entry-level courses by combining analysis of graphs and contextual scientific knowledge. Furthermore, this example shows that a combination of graphical representations and contextual knowledge can serve as a good basis when introducing mediation models using formal inferential statistics afterwards.

6.3. Analysing the influence of dichotomous variables

In contrast to continuous variables, the influence of dichotomous variables can also be analysed using TinkerPlots. In this task, the association of the precipitation in Graz with the PM10 concentration will be analysed. In a first step, a new plot with a continuous PM10 x-axis is created in TinkerPlots. Using the vertical stack option, the distribution of the PM10 values can
be displayed. When clicking the column precipitation, the data dots appear in either blue (there was precipitation on that day) or red (there was no precipitation on that day) as shown in figure 7.

This graph represents the effect precipitation has on the PM10 concentration: Days with lower PM10 concentration are more frequently days with precipitation. To gain further insights into this phenomenon, one can separate the distributions by precipitation. This leads to two different distributions of PM10 concentration: one with precipitation and a second without precipitation. With the option divide, the percentage of days with a PM10 concentration higher than for example 50 μg m⁻³ can be shown. As represented in figure 8, for days with precipitation not only the mean PM10 concentration is lower, but also the percentage of days that exceed the 50 μg m⁻³ mark is much lower (20% of days without precipitation, 6% of days with precipitation).

This relation is even more obvious when the two distributions are compared with boxplots (see figure 9).

6.4. Analysing interaction effects

TinkerPlots is also a good tool to visualize interaction effects. Although regression lines cannot be directly generated with TinkerPlots, there are some options to compare different scatterplots. In this example, the difference between the station with high traffic and the station with low traffic is not displayed by colour-coded dots in a graph, but by separating the data for the two stations into two plots, which can be displayed simultaneously. Figure 10 shows two graphs, the left graph for the station with high traffic, the right graph for the station...
Figure 8. Distributions of average PM10 concentration, divided by precipitation.

Figure 9. Distributions of average PM10 concentration, divided by precipitation and with box-plots.

Figure 10. Scatterplots with a line of best fit for the station with high traffic (left) and low traffic (right).
with low traffic. In a next step, the scatterplots can be compared. There are two options to do so: firstly, a line of best fit can be drawn by hand or secondly, TinkerPlots can generate a ‘Diagonal Reference Line’. By adjusting the ‘Diagonal Reference Lines’ in such a way that it serves as an intuitive line of best fit, it is possible to extract the slope of the straight line (which is displayed when using the software).

Doing so, one can identify that not only the intercept is higher in the graph for the station with high traffic, but also the slope is comparably steeper. Based on these example findings of the exploratory data analysis using TinkerPlots, multiple regression models are calculated. They are reported in the next section.

7. Analysis of the factors influencing PM10 concentration in Graz using multiple regression analysis

The first multiple regression model presented yields an overview of the relevant factors influencing particulate matter concentration in Graz. A multiple regression model for \( \sqrt{PM10} \) was calculated to see which covariates of the dataset have a significant impact on the PM10 concentration. From a scientific point of view, analysis of the PM10 concentration would have been more meaningful, however the square-root transformation used in the model was performed to assure a constant error variance in order to apply multiple regression analysis but also to simplify the data for the students. Additionally, when running a model with PM10 as the dependent variable, the significant predictors stay the same.

A multiple linear regression model was calculated to predict \( \sqrt{PM10} \) based on the precipitation, average wind speed, temperature gradient, PM10 concentration of the day before, whether it is a workday and whether the day is in February or March:

\[
\sqrt{PM10} = \beta_0 + \sum \beta_i X_i + \epsilon; \epsilon \sim N(0, \sigma^2).
\]

A significant regression was found \( (F(7,946) = 318.8, p < 0.000) \) with an \( R^2 \) of 0.700 and every included covariate had a significant impact (table 2).

As one can see in table 2, the \( \sqrt{PM10} \) concentration of the day before has the greatest impact on \( \sqrt{PM10} \) with a beta-coefficient of 0.44, which is not surprising. In addition, \( \sqrt{PM10} \) concentration decreased by \(-0.19\) when there was precipitation, by \(-0.30\) for each increase in m s\(^{-1}\) of average wind speed and increased by \(0.26\) for each increase in °C of average temperature gradient.

For diagnostics, the match between theoretical and observed values was sufficient (figure 11) and the assumption of normal distribution of the studentised residuals was not violated (figure 12).

At this point, it is important to note that the model presented does not claim to be the best possible model for predicting PM10 concentration in Graz, but it serves our educational purposes best. There are several ways of improvement: Adding a lag to the work/holiday variable would improve the model and the indicator for inversion is very conservative since we used the daily average. Additionally, because we used a square-root transformation, the prediction of PM10 would be given by \( \hat{PM10} = (\sum \hat{\beta}_i X_i)^2 \), where \( \hat{\beta}_i \) are the estimates of \( \beta_i \) in the regression model. Nevertheless, the proposed model is sufficient to build the basis for the learning environment.

As shown in table 2, apart from PM10 concentration of the day before, average wind speed has the highest impact on PM10. However, high wind speed is also one of the main

1 For our data sample, February and March are months with especially high PM10 concentration in Graz.
factors preventing inversion. Hence, a possible line of investigation for the students engaging in our learning environment is to test a mediation model. In this example, the mediation model was tested in three steps to check whether the temperature gradient mediates the effect of wind speed on the PM10 concentration. In the first step, it was shown that average wind speed significantly predicts PM10 concentration (see Table 3, Model 1). In the second step, a model was calculated to show that average wind speed also significantly predicts average temperature gradient (see Table 3, Model 2). In the last and third step, a multiple regression model was calculated with average wind speed and temperature gradient as predictors (Model 3 in Table 3). As shown in Table 3, the association of average wind speed with PM10 concentration is reduced when the average temperature gradient is added to the model. These findings support at least partial mediation.

8. Feedback, evaluation and outlook

In a first step, we piloted the described learning environment with 34 physics teacher students. After the teaching sequence, all students filled in a feedback questionnaire. All students
appreciated the use of TinkerPlots and especially highlighted that the software was very intuitive and easy to handle. Regarding the question ‘What did you enjoy the most while working on this assignment,’ more than half of the students highlighted working with authentic data from a real-world context.

While conducting the exploratory data analysis, students also filled in a pre-structured, very detailed report. We are currently analysing these reports with respect to the sequencing according to the PPDAC-cycle and aspects of informal inferential reasoning. First results show that students approached the assignment in quite different ways: Some students only formulate one research question, which they investigate in multiple ways using multiple representations of the same relationship in order to improve their conclusion. Some students develop several research questions, while basing their conclusion on only one graphical representation. These are the first results, which need a more detailed qualitative and quantitative analysis. While this article aims to introduce the concept of this learning environment, detailed analysis of the student reports will be covered elsewhere.

9. Conclusion

Students enrolled in a science or technology-related study should not only have the possibility to conduct experimental investigations, but also other types of scientific inquiry. In this article, we present a learning environment that introduces informal inferential reasoning and exploratory data analysis as a prerequisite before introducing multiple regression analysis. Doing so, students should not only learn about the mathematical underpinning, but also understand mind-sets that are involved when conducting correlational investigations in science. In the exploratory data analysis part of our learning environment we are using data from online data repositories. As software we chose TinkerPlots since it eliminates time for routine tasks like calculations or further preparation of data in order to allow more time for higher order thinking.
Table 3. Summary of three regression models used to show the mediation effect of temperature gradient when regressing average wind speed on \(\sqrt{PM_{10}}; p < 0.001^{***}\).

| Model 1—\(PM_{10}\) as dependent variable | Model 2—temperature gradient as dependent variable | Model 3—\(\sqrt{PM_{10}}\) as dependent variable |
|------------------------------------------|-----------------------------------------------|----------------------------------------------|
| B | SE B | \(\beta\) | B | SE B | \(\beta\) | B | SE B | \(\beta\) |
|------------------------------------------|-----------------------------------------------|----------------------------------------------|
| Constant | 6.62 | 0.08 | -1.65 | 0.20 | 6.92 | 0.07 |
| Wind speed | -1.90 | 0.10 | -4.05 | 0.26 | -0.45 | 0.26 |
| Temperature gradient | 0.18 | 0.01 | 0.18 | 0.01 | 0.46 | 0.05 |
| Adjusted \(R^2\) | 0.28 | 0.20 | 0.20 | 0.20 | 0.45 | 0.45 |
| \(F\)-statistics | \(F(1952) = 372.7^{***}\) | \(F(1952) = 244.6^{***}\) | \(F(2951) = 387.6^{***}\) |
Overall it has been shown that the use of graphical representations made with TinkerPlots is a good basis for the design of learning environments concerning the exploratory analysis of data. Furthermore, the described multiple regression models of $PM_{10}$ are sufficient to describe the relevant factors influencing PM10 concentration in Graz. However, when using graphical representations for analysis it is important to emphasize the integration of contextual knowledge in the data analysis process. In addition, a meaningful formulation of conclusions considering limiting factors is of importance. The idea of this article can be used in a wide set of areas, for example, the analysis of time series in correlational studies or data from experiments can be analysed with TinkerPlots as well. Additionally, many more multiple regression-related aspects can be discussed with our setting. For example, the concept of heteroscedasticity and the need for data transformations can be discussed with our data sample as well.

Acknowledgments

The authors acknowledge the financial support by the University of Graz.

ORCID iDs

Thomas Schubatzky @ https://orcid.org/0000-0002-0736-7468
Claudia Haagen-Schützenhöfer @ https://orcid.org/0000-0002-6245-7888

References

Chance B, Ben-Zvi D and Garfield J 2007 The role of technology in improving student learning of statistics Technol. Innov. Statist. Educ. 1 1
Esworthy R 2013 Air quality: EPA’s 2013 changes to the particulate matter (PM) standard Congressional Research Service R42934 Library of Congress
Finzer W 2007 Fathom Dynamic Data Software (Emeryville, CA: Key Curriculum Press)
Frischemeier D and Biehler R 2013 Design and exploratory evaluation of a learning trajectory leading to do randomization tests facilitated by TinkerPlots Proc. 8th Congress of European Research in Mathematics Education pp 798–808
Furtak E M, Seidel T, Iverson H and Briggs D C 2012 Experimental and quasi-experimental studies of inquiry-based science teaching Rev. Educ. Res. 82 300–29
Hörmann S, Pfeiler B and Stadlober E 2005 Analysis and prediction of particulate matter PM10 for the winter season in Graz Austrian J. Stat. 34 307–26
Kim K-H, Kabir E and Kabir S 2015 A review on the human health impact of airborne particulate matter Environ. Int. 74 136–43
Konold C 2007 Designing a data analysis tool for learners ed P Lovett and M Shah Thinking with data: The 33rd Annual Carnegie Symp. on Cognition pp 267–91 (NJ: Lawrence Erlbaum Associates)
Konold C and Miller C D 2005 TinkerPlots: Dynamic data exploration: Computer Software (Emeryville, CA: Key Curriculum Press)
Lederman J S, Lederman N G, Bartos S A, Bartels S L, Meyer A A and Schwartz R S 2014 Meaningful assessment of learners’ understandings about scientific inquiry-The views about scientific inquiry (VASI) questionnaire J. Res. Sci. Teach. 51 65–83
Lederman N G, Lederman J S and Antink A 2013 Nature of science and scientific inquiry as contexts for the learning of science and achievement of scientific literacy Int. J. Educ. Math., Sci. Technol. 1 138–47 https://ijemst.net/index.php/ijemst/article/download/19/19
Mastrandrea M D, Mach K J, Plattner G-K, Edenhofer O, Stocker T F, Field C B, Ebi K L and Matschoss P R 2010 The IPCC AR5 guidance note on consistent treatment of uncertainties: a common approach across the working groups Climatic Change 108 675–91
Schneider J, Spangl W, Placer K and Moosmann L 2005 Abschätzung der Gesundheitsauswirkungen von Schwebestaub in Österreich (Vienna: Umweltbundesamt)
Watson J and Donne J 2009 TinkerPlots as a research tool to explore student understanding Technol. Innov. Stat. Educ. 3 3
WHO 2013 Health effects of particulate matter ed J Q Koenig Health Effects of Ambient air Pollution: How Safe is the Air we Breathe? pp 115–37 (Berlin: Springer)
Wild C J and Pfannkuch M 1999 Statistical thinking in empirical enquiry Int. Stat. Rev. 67 223–65
Wild C J, Utts J M and Horton N J 2018 What is statistics ed D Ben-Zvi et al International Handbook of Research in Statistics Education (Berlin: Springer)