Kazak, Fujita and Turmo, Mathematical Thinking and Learning, 2021

Students' informal statistical inferences through data modeling with a large multivariate dataset

Sibel Kazak[a], Taro Fujita[b], and Manoli Pifarre Turmo[c]

aDepartment of Mathematics and Science Education, Pamukkale University, Denizli, Turkey; bGraduate School of

Education, University of Exeter, Exeter, UK; cDepartment of Psychology and Pedagogy, University of Lleida, Lleida, 5

Spain

In today's age of information, data are very powerful in making informed decisions. Data analytics is a field that is interested in identifying and interpreting trends and patterns within big data to make data-driven decisions. We focus on informal statistical inference and data modeling as a means of developing students' data analytics skills in school. In this study, we examine how students apply the data modelling process through measuring and visualizing variability to make inferences from a real dataset when exploring the data to identify trends, patterns and possible relationships between variables using technological tools, such as CODAP and Excel. We analyzed 17-18-year-old students' written reports on their explorations of data supplied by third parties. Students used a variety of statistical measures and visualizations to account for variability in analyzing data. Context and uncertainty played a significant role in making inferences and predictions beyond the data.

Keywords: data analytics; data modeling; informal statistical inference; upper secondary

Introduction

Over the past three decades, there has been a growing acceptance of statistics as part of mathematics education around the world. Overall, the statistics strand of the school

mathematics curricula in many countries (e.g., Common Core State Standards Initiative, 2010; Department for Education, 2014) aims to develop students' understanding of basic concepts, tools, and procedures in statistics and probability, and their statistical problem solving skills (e.g., formulating a question, collecting, organizing and analyzing data, and interpreting results). In everyday life, thinking statistically is essential to make better decisions in situations under uncertainty, on the basis of data. This type of thinking requires being able to use fundamental ideas of statistics, to understand and use the context relevant to the problem being solved, and to critique and evaluate the results from the problem solved (Ben-Zvi & Garfield, 2004).

In the practice of statistics, modeling and reasoning with statistical models are considered as essential components of statistical thinking (Wild & Pfannkuch, 1999). Similar to the notion of mathematical modeling described by Lesh and Doerr (2003), statistical modeling refers to the process in which modelers develop and use some conceptual systems to represent, interpret, explain and make sense of a problem situation by thinking about variability and chance. The context of the real-world problem from which data come is also an essential component of this modeling process. With the availability of technological tools, such as TinkerPlots (Konold & Miller, 2011), statistical modeling emerged as a possible way to help learners at different ages to develop an understanding of big statistical ideas (e.g., distribution, variability, sample and sampling, uncertainty, and informal statistical inference) as well as to connect data, chance and context (see special issues Statistics Education Research Journal, 16(2), 2017; ZDM 50(7), 2018).

In these efforts, statistical modeling is interpreted and utilized in various ways, including: 1) engaging students in the modeling process for developing their understanding of distribution, statistical measures and data representations as generalized

models for making inferences (e.g., Büscher & Schnell, 2017; Doerr, delMas, & Makar, 2017; Fielding-Wells, 2018; Makar & Allmond, 2018) and 2) engaging students in building models of real data using random generating devices available in TinkerPlots to produce outcomes (simulated data) resembling the real data distribution (e.g., Ainley & Pratt, 2017; Aridor & Ben-Zvi, 2018; Patel & Pfannkuch, 2018). The latter approach involves creating models using chance ideas to represent an uncertain phenomenon. In the former approach, known as data modeling (Lehrer & English, 2018), the process of modeling takes place when students deal with variability in the context of a problem and use statistical concepts and tools for organizing, structuring, measuring and representing data. These concepts and tools become a model for drawing inferences from data. While the existing research (Büscher & Schnell, 2017; Doerr, delMas, & Makar, 2017; English & Watson, 2018; Fielding-Wells, 2018; Lehrer & Schauble, 2004; Makar & Allmond, 2018) tended to study young students' data modeling with small samples, further research can provide insights into students' data modeling process through engaging students in working with larger multivariate datasets.

As pointed out by Ridgway (2016), statistics curriculum for schools in the UK and other countries mostly involve drawing conclusions from a small sample to make generalizations about a population. With the new developments in digital and data technologies, huge quantities of data are generated from a variety of sources every day and we need people with skills to turn this large volume of structured or unstructured data into insight and action to improve service delivery, policy making, quality of life and so on. This paper reports on a study that aims to develop such skills for 17-18-year-old students as part of a larger project. In this study, we focus on students' data modeling when solving a real-world problem in the context of air pollution by exploring a large multivariate dataset, supplied by third-party providers, with the use of technological tools,

such as Excel and CODAP. Our research question is: How do students measure and visualize variability to draw informal inferences when exploring trends/patterns and relationships in a large multivariate dataset with the use of technological tools?

The context of the study

In today's age of information, emerging data sources provide large, rich datasets, and such data is very powerful in making informed decisions to influence policy, public opinion and business practices (Ridgway, 2016). Data analytics, which is a relatively new field, refers to processes used in identifying or discovering trends and patterns inherent in the data to provide useful insights for making data-driven decisions (Piccano, 2012). The project X aims at developing such skills around 'data analytics' for students at various ages from 9 to 18 using an innovative and student-centered approach in partnering schools.

In this project, data analytics (DA) is conceptualized as a process of *'engaging creatively in exploring data, including big data, to understand our world better, to draw conclusions, to make decisions and predictions, and to critically evaluate present/future courses of actions*'. Although DA is somewhat a new concept in schools, it can be built on Wild and Pfannkuch's (1999) description of the statistical investigative cycle, called the PPDAC (Problem, Plan, Data, Analysis, Conclusion), which often forms the basis of statistical investigation taught in schools in many countries, such as in New Zealand (Arnold, 2013). In the PPDAC cycle, the first step is to understand and define the problem (*Problem*); the second step is to decide appropriate measures and variables (*Plan*), and the third step is to gather and clean data (*Data*). As seen in the definition of DA above, data analytics usually begin with '*exploring data'* which is at halfway through the third step (*Data*) of PPDAC cycle. However, "to do a good job, [a data] analyst needs to develop a good understanding of what has gone on before, particularly how the data was

actually obtained" (Wild, 2018, p. 1). So guiding students through the *Problem*, *Plan* and *Data* stages of the PPDAC cycle can help them engage in '*exploring data'* in DA. The next two stages of the PPDAC cycle involve using appropriate data visualization tools and calculations to look for patterns and relationships (*Analysis*) and making interpretations and inferences and communicating findings (*Conclusion*). In DA, these processes are required to '*draw conclusions'*, to '*make decisions and predictions'*, and to offer new ideas to '*take courses of actions'*.

In such statistical problem-solving processes, the role of variability is salient. For example, the Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report (Franklin et al., 2007) for pre-K to 12 level explicitly emphasize the role of variability for each component of the statistical investigative process: 1) 'anticipating variability' in formulating the question − writing a statistical question entails an anticipation of an answer based on data that vary; 2) 'acknowledging variability' in designing the data collection − planning the method of data collection requires an acknowledgement of variability in data with an intent to reduce it; 3) 'accounting of variability' in statistical analysis − analyzing data involves accounting for variability by using distributions; 4) 'allowing for variability' in interpreting results − making a generalization beyond the data needs to allow for variability in the data. Given that "statistical models are developed as accounts of variability" (Lehrer & English, 2018, p. 230), Lehrer and English view data modeling as a way to promote students' understanding of variability. Lehrer and Schauble (2004) describe data modeling as a cyclic process,

> in which understanding the problem at hand is prerequisite for beginning to think about potential feasible solutions. Decisions about what and how to measure, the ways that data should be structured and displayed, and the conduct of inference are always rooted firmly in knowledge about the world. (p. 636)

Since the role of variability is essential in the whole DA process and decision making is one of the key components of DA, we focus on data modeling and informal statistical inference to help students develop key understandings in statistics and competencies relevant to DA.

Theoretical Background and Literature Review

In this section, firstly we elaborate on a framework for DA and its connection to other frameworks in the statistics education literature. Then, we focus on data modeling as a broad lens to promote DA skills for exploring large datasets by using data visualization tools and appropriate statistical calculations, and drawing inferences. Finally, in relation to the data modeling framework we review existing research on specific statistical topics, namely variability and inference.

Data Analytics (DA) Framework

Our conceptual framework for DA in schools has two components: data analytics cycle and competence areas (Figure 1). At the heart of this framework is the cyclic process of acts that we expect when students engage in DA. It is called ‘Data Analytics (DA) Cycle’ drawn on PPDAC statistical inquiry cycle (Wild & Pfannkuch, 1999), statistical thinking process (Wild, Utts, & Horton, 2011) and informal statistical inference (Makar & Rubin, 2009, 2018).

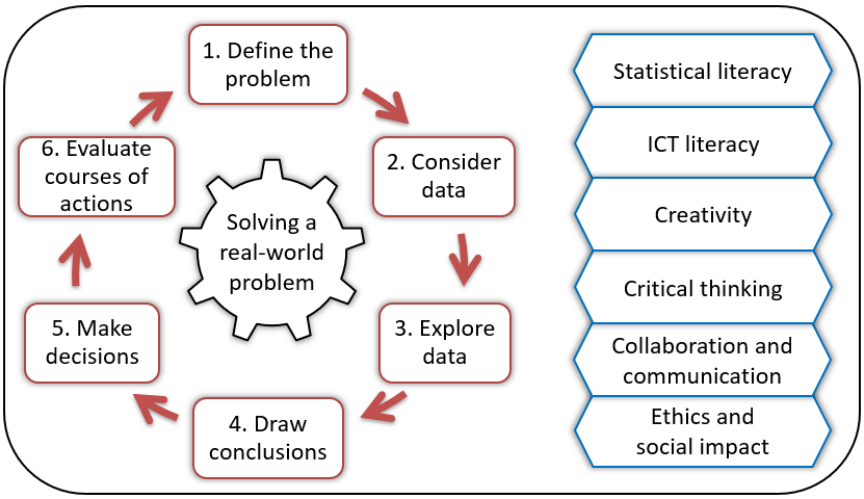


Figure 1. Data analytics cycle and competence areas in DA framework

In the context of project X, this investigative cycle is particularly concerned with solving real-world problems related to weather, which affects a huge variety of natural phenomenon and human behavior. Since DA is associated with data-driven decision making, in order to figure out what actions to take for solving a problem using data, one needs to acquire more knowledge (Wild et al., 2011) and gain a better understanding of data as seen in the first three stages of the PPDAC cycle (Problem, Plan, Data). Hence, the DA cycle starts with (1) *defining the problem* with recognition of the need for data and formulation of specific questions to be answered using data. Next, (2) *considering data* involves deciding what individuals or entities to obtain data on, what to measure and how to collect them, as well as collecting and tidying data. Then (3) *exploring data* comprises analyzing data using data visualization tools, appropriate calculations and statistical models. Similar to the last two stages of the PPDAC cycle (Analysis, Conclusion), this leads to (4) *drawing conclusions* and (5) *making decisions* by using data as evidence for generalizations beyond describing the given data and making predictions with an articulation of uncertainty. Furthermore, these prior steps guide (6) e*valuating courses of actions* in connection with the problem defined earlier − What actions need to be taken? (e.g., collect more data, do more analyses, ask experts and so on).

Moreover, this DA cycle is complemented with various competence areas that are in line with "Framework for 21st Century Learning" (http://www.p21.org/about-us/p21-framework) by the Partnership for the 21st Century Learning (P21) as well as The Royal Society's (2016) report (in partnership with the Royal Statistical Society) on the need for data analytics skills. These include:

- *Statistical literacy*. (a) Understanding basic statistical concepts, vocabulary, procedures and techniques; (b) Interpreting and evaluating statistical information or data-based claims where they are contextualized; (c) Communicating opinions about the statistical information and concerns about the soundness of statistical arguments (Gal, 2002; Garfield, delMas and Chance, 2003).
- *Information and Communication Technologies (ICT) literacy*. Using technology or computing capabilities to understand and solve problems; to visualize, model, code and organize data; and to communicate statistical information.
- *Critical thinking*. (a) Reasoning effectively; (b) Analyzing and evaluating data-based evidence and arguments; (c) Interpreting and making conclusions based on the best analysis; (d) Reflecting critically on processes in solving problems.
- *Creativity*. (a) Using different approaches/techniques in a particular task; (b) Generating new ideas and methods; (c) Being open to new diverse perspectives.
- *Communication and collaboration*. (a) Using available tools and language effectively in articulating thoughts/ideas in the problem context; (b) Working with others effectively in groups.
- *Ethics and social impact*. (a) Taking responsibility to both act and refrain from certain actions with the interests of society at large in mind; (b) Demonstrating consciousness about the challenges in the digital age.

Data modeling

Among various interpretations of statistical modeling in the statistics education literature (see Pfannkuch, Ben-Zvi & Budgett, 2018), data modeling is seen as accessible especially to young students in conducting statistical modeling with considerations about variability and context (e.g., English, 2010; Kazak, Pratt & Gökce, 2018; Shinohara & Lehrer, 2017). According to Lehrer and English (2018), "representations of variability take many forms, including visual display, data structures, measures, and models" (p. 233) and putting together these various forms to get a sense of variability in data when making inferences is considered as data modeling.

Based on prior work of Lehrer and Schauble (2004), Lehrer and English (2018) outlined the components of data modeling as seen in Figure 2. Part I of this figure involves mutually dependent (shown with double arrows) conceptions of designing and conducting a statistical investigation, such as posing questions, selecting and measuring attributes, and generating a sample. Part II of Figure 2 refers to the process of modeling variability in connection with other mutually dependent forms of activities, such as organizing, structuring, measuring, and representing data, which lead to making inferences in relation to the questions investigated. In this study, we focus on the lower portion, and more specifically, the following key ideas as described by Lehrer and English (2018):

- Modeling variability requires reasoning about data as aggregate rather than focusing on values of individual cases. Visualizing variability (i.e., use of representations showing trends of variability in the data) and structuring variability by measuring (i.e., use of statistical measures of center and variation as indicators of distribution features) are the two approaches to modeling variability.
- Making inferences entails needing to warrant conjectures about statistical questions or claims. So, inference can be seen as noticing characteristics of

displays/measures and using them to justify claims, as well as arguing based on evidence and acknowledging uncertainty.

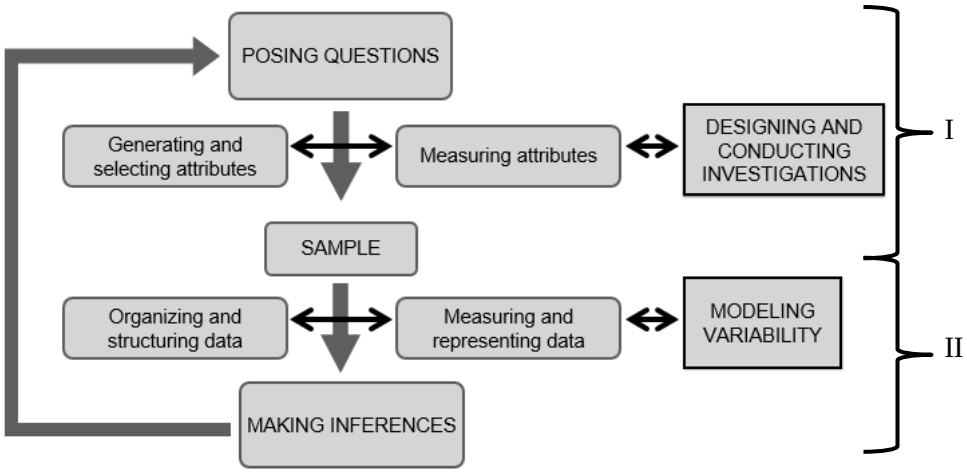


Figure 2. Components of data modeling (Lehrer & English, 2018)

In our research, the DA cycle (Figure 1) offers a gateway to data modeling, which is based on the iterative statistical inquiry cycle (Pfannkuch, et al., 2018). Since measuring and representing data and making inferences entail dealing with variation, modeling, and considering context (ibid., 2018), we view the students' actions in the DA cycle as the data modeling process. Within the context of the larger project, the data modeling approach is utilized as a way of "[g]aining more insight into a particular situation through statistical modeling and learning more about the context with the goal of prediction, explanation or control" (ibid., 2018, p. 3) with combination of engaging students in the DA cycle.

Variability and inference

'Accounting of variability' in analyzing data and 'allowing for variability' in interpreting results (Franklin et al., 2007) are important components of the lower portion of the data modeling framework in Figure 2. Hence, we now turn to prior studies that can offer

insights on how students deal with variability by measuring and visualizing data through the data modeling process.

English and Watson (2018) described a framework for a 'modeling with data' approach with the following four components: "working in shared problem spaces (boundary interactions) between mathematics and statistics; interpreting and reinterpreting problem contexts and questions; interpreting, organising and operating on data in model construction; and drawing informal inferences." (p. 104). They examined 11-year-old students' data modeling process for selecting six swimmers for the national team to compete in the 2016 Olympics based on a real dataset provided. Researchers argued that this approach had potential to foster students' understanding of foundational statistical ideas and processes because such data modeling problems were complex and ill-structured. Therefore, students would need to make sense of problem contexts, organize data so that they could manage and work with the data, and make decisions and inferences beyond the data. For instance, when dealing with the inherent variation in the data provided, students needed to determine what variables to use and define for selecting team members, such as personal best times and consistency in performance. Students also tended to consider problem context in relation to selecting variables during model construction. Moreover, they identified trends in the data and made predictions for swimmers' performance in the future with acknowledgment of uncertainty due to data limitation or chance variation.

Fielding-Wells (2018) examined 10-11-year-old students' use of various data models afforded by TinkerPlots software in making inferences. These models and technological tools offered a conceptual approach to fostering students' explorations of key statistical ideas, such as distribution, variation and center, within meaningful contexts. More specifically, students designed catapult planes using paper, and collected

data on flight distance to find out the best paper size to make a catapult plane. They organized the collected data by constructing multiple representations, such as dot plots and hat plots, in TinkerPlots for two different plane designs. These data models allowed students to identify patterns in the distributions and visualize variability in the actual data. It was also noted that when students drew conclusions based on data, they seemed to begin to use probabilistic language and qualifiers (e.g., "will probably go further" and "most of the time" (p. 1133)) to express their (un)certainty in their statements.

Using an 'emergent modeling approach', Büscher and Schnell (2017) studied 14-year-old students' understanding of statistical measures in relation to average and variation as they developed models from given small datasets to report on the change in the size of the arctic ice area between 1982 and 2012. Within the context of weather and climate, this research showed students' conceptual shift from 'models of a situation' ("context specific and act as a stand-in used for talking about one specific phenomenon") to 'models for reasoning' ("generalised over several situations and allow organising a variety of similar and new contexts accordingly") (Büscher & Schnell, 2017, p. 148). In the former, students used the model to understand the given problem situation, and the statistical measures became tools to structure and underline certain aspects of distributions in connection with their contextual knowledge. In the latter, on the other hand, the measures were used as models for reasoning about distributions, which could be applied to other problem situations. In other words, these applied measures were evaluated in terms of their adequateness in making inferences. It was also argued that development of statistical knowledge and contextual knowledge evolved in relation to each other during the emergent modeling process.

As seen in the above studies, data modeling has been researched mainly with young students to develop the foundations for the formal concepts and ideas needed in

the practice of statistics, such as measures of center and variation, statistical inference, and modeling. We argue that data modeling can also provide insights for looking into opportunities to support the statistical learning of older students who do not study advanced topics in statistics before going to university or workplaces.

Our other focus on the data modeling framework (Figure 2) is making inferences. The concept of inference in statistics refers to "drawing conclusions about populations or processes based on sample data" (Zieffler, Garfield, delMas, & Reading, 2008, p. 40). More specifically, formal statistical inference includes certain techniques, such as hypothesis testing and confidence intervals, to draw conclusions from the sample. On the other hand, data-driven decision making, which is essential in our DA framework, is often referred to as informal statistical inference when students make inferential statements but not necessarily use formal inferential procedures in statistics (Makar & Rubin, 2009). According to Makar and Rubin, this inferential process involves the following components: 1) available data are used as evidence to justify the inference, 2) generalization goes beyond these data and 3) uncertainty is articulated in making a statistical inference. The underlying reasoning that leads to informal statistical inference is called informal inferential reasoning (Makar, Bakker, & Ben-Zvi, 2011).

Starting from the early school years, informal statistical inference has been conceptualized as a holistic approach to develop students' understanding of foundational statistical concepts, such as distribution, center, and variation, (Bakker & Derry, 2011). More recently, informal statistical inference has been studied in relation to statistical measures and displays, uncertainty, context, and modeling (e.g., Ben-Zvi, Aridor, Makar, & Bakker, 2012; Doerr et al., 2017; Fielding-Wells, 2018; Langrall, Nisbet, Mooney, & Jansem, 2011; Makar, 2014; Pfannkuch, 2011).

Students are expected to develop an understanding of statistical measures in the school curriculum. However, earlier research suggests that, despite students' familiarity with the algorithm for the mean, many of them tend to have difficulty in using it as a representative value for a dataset and as an appropriate measure in data analysis (Lavigne & Lajoie 2007; Watson & Moritz, 2000). Makar (2014) illustrated that the use of informal inferential reasoning had great potential in helping 8-year-old children to develop richer conceptions of average in an inquiry-based learning environment while making sense of average in relation to reasonableness, outliers, modal values, comparing groups and inference. In another inquiry-driven classroom-based study on the development of informal inferential reasoning from a modeling perspective, Doerr et al. (2017) used a model development sequence to facilitate 10-11-year-old students develop improved models of uncertain phenomena involving flight times of paper helicopters. During this sequence of activities, students' use of a data representation (dot plot) evolved from being a tool for visualizing variability in the data to a generalizable model for comparing two sets of data to draw inferences based on evidence.

Ben-Zvi et al. (2011) emphasized that inferential statements from a sample to a larger population should include an articulation of uncertainty because these claims go beyond the data at hand. They researched two groups of 2-3 students (aged 10-11) during their investigations of student survey data from a very small sample (n=8) and then the whole class data (n=27) by using TinkerPlots software. When students made inferential predictions for the whole class from a very small sample, they tended to be extremely confident in their claims or express absolute uncertainty (e.g., nothing can be concluded). In the next investigation with a larger sample, students' inferences from their 5th grade class to a bigger group (all 5th graders) showed an emergence of using probabilistic language in their statements, such as "the chances are … really small", and qualitative

expressions of how certain they felt about the inference, e.g., "it seems that…." (Ben-Zvi et al., 2011, p. 923). It was argued that this emerging articulation of uncertainty was facilitated by introducing students to increasing sample sizes selected from the same population combined with "what if" questions about whether their inferences would apply to a larger sample.

Others reported on the role of using context knowledge when students were engaged in making inferences based on given authentic data (Langrall, Nisbet, Mooney, & Jansem, 2011; Pfannkuch, 2011). These studies indicated a variety of ways students used context knowledge in their data explorations. For example, Langrall et al. (2011) state that context expertise promotes the use of context knowledge to develop a new insight to the task, to identify useful data, to provide justifications for the inferences made and to foster the recognition of the limitations of a claim. In particular, the inferential explanations were grounded in the knowledge of context that was used as a story behind the data as well as enabling students to go beyond the data to make inferences. Pfannkuch (2011) suggests that, not only the knowledge of data-context, but also learning-experience-contexts can play an important role in developing students' inferential reasoning. For instance, prior statistical knowledge of students and teacher, student-teacher interactions and task sequence can mediate learner's informal statistical inference process in a learning environment as parts of learning-experience-contexts.

Hence, we argue that promoting informal statistical inference, defined as "decision-making in relation to a statistical question for a population based on evidence from a sample and acknowledging a degree of uncertainty in that decision" (Watson & English, 2018, p. 36), can support the skills needed for the DA cycle.

Methodology

This work is part of project X, which aims to innovate and extend best practice in the teaching of DA through student-centered, project-based learning, focusing on the impacts of weather. The end product is an on-line resource for schools to develop their DA projects building on our examination of the state-of-the-art findings emerging from the pilot projects conducted in collaboration with teachers in ten partnering schools (from primary to upper secondary) in three European countries. To this purpose, a design study method (Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003) is used for developing, testing and revising conjectures about how students develop skills and competencies related to DA and instructional materials to support it with the use of technology tools, such as Excel and CODAP (Common Online Data Analysis Platform, https://codap.concord.org/). CODAP is a free web-based data visualization tool and provides opportunities "to move evidence to the centre of the curriculum, and to encourage engagement from a variety of disciplinary perspectives with statistical thinking" (Ridgway et al., 2017, p. 4). In this paper, we focus on one study from a UK partnering school to investigate how students measure and visualize variability to make inferences from a large multivariate dataset supplied by third-party providers by using technological tools, such as CODAP and Excel.

Participants and context

The participants were 42 17-18-year-old students (mostly males) studying for the Business and Technology Education Council (BTEC) qualifications, which are for both university entries and employers, in southwest England. In fall 2018, the students took 'Mathematics for IT Practitioners', a subsidiary subject to their main subject area (Informational Technology-IT). In 'Mathematics for IT Practitioners', 9 hours were allocated for 'interpretation of data', aiming at developing IT students' statistical literacy through using statistical measures, e.g., mean, median, standard deviation, and creating

charts from data supplied by third parties. The students have studied basic statistical concepts, such as appropriate graphical representation involving discrete, continuous and grouped data; and appropriate measures of central tendency (mean, mode, median) and spread (range, consideration of outliers), scatterplots, etc. in lower secondary schools.

Task

This study focuses on the pilot project aimed to investigate airborne pollutants, such as particulate matter (PM10) and Nitrogen Dioxide (NO_2). These pollutants have been identified as being dangerous to health and are generated by industrial and consumer activities, specifically diesel vehicle emissions. During the project, students were encouraged to explore real datasets supplied by third parties to identify and interpret trends and patterns in the pollution data.

Extracted from a larger dataset (approx. 28000 records) containing anomalies (missing data, incorrect data etc.) provided by City Council, one of the datasets given to students included 720 cases with hourly PM10 data from November 2015, 2016 and 2017 and time of the day (morning, afternoon, evening, night) in two different locations in the city. Figure 3 shows part of the dataset used for investigation.

1	Date/Time	TimeOfDay	RAMM 2017	RAMM 2016	RAMM 2015	Alphington 2017	Alphington 2016	Alphington 2015
2	11.1.2016 00:00	Night	25,336	20,883	30,212	20,203	20,848	31,964
3	11.1.2016 01:00	Night	20,333	18,753		18,925	19,486	
4	11.1.2016 02:00	Night	18,313	16,794	26,339	17,75	16,773	38,714
5	11.1.2016 03:00	Night	17,825	19,164	22,216	18,386	18,452	23,825
6	11.1.2016 04:00	Night	15,978	19,598	16,904	17,08		17,782
7	11.1.2016 05:00	Night	16,754	19,734	23,793	17,786	19,281	16,528
8	11.1.2016 06:00	Night	19,835	28,332	27,175	20,57	20,869	19,669
9	11.1.2016 07:00	Morning	30,277	37,86	24,357	29,174	24,797	15,204
10	11.1.2016 08:00	Morning	38,563	41,492	29,837	38,337		19,746
11	11.1.2016 09:00	Morning	50,066	50,261		39,064	30,547	
12	11.1.2016 10:00	Morning	39,171	28,161		55,075	24,253	
13	11.1.2016 11:00	Morning	24,68	26,708	30,08	26,393	25,757	26,865
14	11.1.2016 12:00	Morning	23,293	32,163		25,152	25,866	
15	11.1.2016 13:00	Afternoon	24,168	51,161	20,564	21,905	31,235	22,327
16	11.1.2016 14:00	Afternoon	24,853	40,301	22,481	20,727	39,602	21,406

Figure 3. Example of the dataset including November PM10 data in two locations for 2015-17

Table 1 shows various tasks students engaged in according to the DA framework (Figure 1). In the first two phases of the DA cycle, students were encouraged to think about data collection methods and the reliability and accuracy of the data supplied to them by third parties, and to consider the questions or hypotheses they would like to answer using the data. In the third phase, students were expected to use statistical measures and data visualization tools, such as mean, median, mode, range, interquartile range, standard deviation, histogram, box plot, stem and leaf plot, correlation and so on, to explore data. As part of the ICT literacy competence in the DA framework, CODAP was used for analysis and investigation of data, but they were free to use other tools, such as Excel. While critical thinking competence was needed in drawing conclusions and making decisions based on the sample data, group work, sharing ideas and presenting results in the class were anticipated as part of the communication and collaboration competence. They worked (individually/in small groups) for a total of 9 hours during the project.

Table 1. Tasks involving the DA cycle and DA competencies.

DA Cycle	DA Competencies
1. Define the problem ● Students work in groups to identify a question or hypothesis to test.	**Statistical literacy** ● Mean, median, mode ● Range, IQR, Standard Deviation ● Histogram, Box plot, Stem & Leaf ● Interpreting data ● Correlation **ICT literacy** ● Excel, CODAP **Critical thinking**
2. Consider data ● Students explore the datasets to identify its validity and accuracy.	
3. Explore data ● Students import data into CODAP and use various tools to explore relationships (correlation), spread and statistical measures.	

4. Draw conclusions ● Students attempt to draw meaningful conclusions from their analysis to test their hypothesis or answer their question.	● Drawing conclusions ● Making decisions **Communication and collaboration** ● Group working ● Sharing ideas ● Presenting results
5. Make decisions ● Where possible, students propose solutions to the question they have posed based on their analysis.	
6. Evaluate courses of actions ● Students evaluate the methods they have used to carry out their analysis and the results they have obtained. Students conclude by writing a report on their project individually.	

After relatively free explorations with the data, the students were asked to write a report on the questions which were semi-structured: Q1) 'Have PM10 levels reached a dangerous level?', Q2) 'Are PM10 levels rising over time?', Q3) 'What time of day are PM10 levels the highest?' and Q4) 'Is there any correlation between PM10 levels at the two sites?'. These questions were suggested by the teacher because during free exploration with the PM10 level data, students often lost direction of their investigations and needed more guidance.

Data analysis

After a random sample of student work was selected by the teacher, ten students' written responses (each 3-10-page documents) to the semi-structured questions posed by the teacher were used for our qualitative analysis. Using progressive focusing (Parlett & Hamilton 1972), we analyzed students' written work in relation to data modeling components, i.e., measuring and visualizing variability and making inferences, within the third and fourth phase of the DA cycle ('explore data' and 'draw conclusions'). First, all student responses were organized in a table by the visualization tools and statistical measures used and inferential statements written for each investigation question. Then,

through multiple stages of analysis of each student's work conducted independently by the researchers, emerging insights developed from the following foci: 1) how students measured and visualized variability when exploring trends/patterns in the given real data with air pollution context, 2) what visualization tools, concepts (such as measures of central tendency and variation) and strategies were used in making inferences with considerations of variability and context, and 3) how they expressed the informal inferences drawn with considerations of uncertainty. Next, researchers compared and discussed their analyses to interpret these foci.

Results

In reporting the results from students' written work, we focus on 'measuring and visualizing variability' and 'informal inferences' as part of the data modeling process within the third and fourth phases of the DA cycle ('explore data' and 'draw conclusions'). To answer our research question, we describe how students measure and visualize variability in the data, including November PM10 values in two locations for 2015-17, to draw inferences when reporting on the following questions: Q1) 'Have PM10 levels reached a dangerous level?', Q2) 'Are PM10 levels rising over time?', Q3) 'What time of day are PM10 levels the highest?' and Q4) 'Is there any correlation between PM10 levels at the two sites?'. In the subsequent sections, we use the following keywords Q1-'Level of danger', Q2-'Rising overtime', Q3-'Time of day' and Q4-'Correlation' in reference to these questions and S1, S2, etc. to refer to Student #1, Student #2 and so on.

Students' ways of measuring and visualizing variability to identify patterns and trends

According to the data modeling framework (Lehrer & English, 2018), modeling variability entails measuring and visualizing variability with the use of statistical measures and data representations within the 'explore data' phase of the DA cycle. In the

given dataset (see Figure 3), variation occurs in the PM10 levels by hours over 30 days in November across years 2015, 2016, and 2017 in two different locations in the city. Thus, interpretation of data requires a consideration of variation within and across each location and year. To be able to answer the question, 'Have PM10 levels reached a dangerous level?' [Q1-'Level of danger'], students also need to think about the meanings of "dangerous level" and "reaching a dangerous level" for airborne particulate matter (PM10). So at the beginning of their reports, they shared the safe and dangerous levels of PM10 from various official sources as a reference. While most students used the table from a government website, others referenced the World Health Organization and the European Union standards (Figure 4). As seen in these examples, setting a safe limit also varies depending on the source organization.

Index	1	2	3	4	5	6	7	8	9	10
Band	Low	Low	Low	Moderate	Moderate	Moderate	High	High	High	Very High
µg/m³	0-16	17-33	34-50	51-58	59-66	67-75	76-83	84-91	92-100	101 or more

(The source: UK Air website, Department for Environment Food & Rural Affairs)

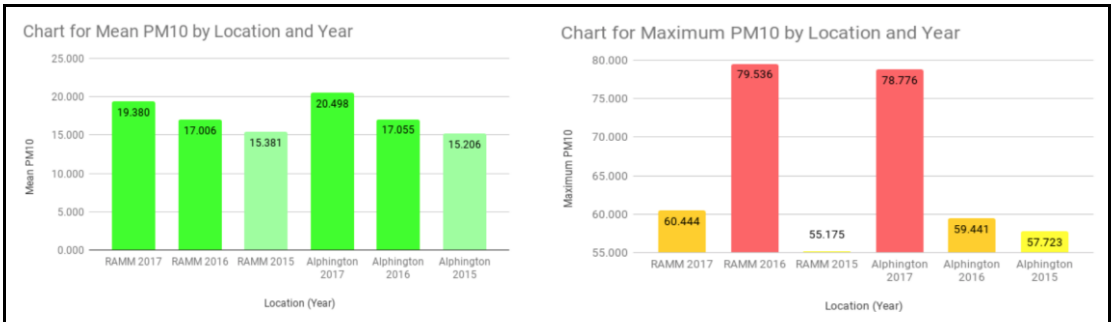
"20 µg/m^3" (The source: The World Health Organization)

"40 µg/m^3" (The source: The European Union standards)

Figure 4. PM10 safe levels from various sources used by the students

When exploring trends and patterns in the given real data in the 'Explore data' phase of the DA cycle, all but one student tended to use statistical measures for center, such as mean, median, and quartiles, and one of them also considered a measure of spread, such as standard deviation, to answer the first two questions ('Have PM10 levels reached a dangerous level?'[Q1-Level of danger], 'Are PM10 levels rising over time?'[Q2-'Rising overtime']). Only four of them considered data representations, such as a bar chart, dot plot, or line graph, to visualize the patterns and trends in the data for these two questions. Others used statistical measures arranged in a table. For instance, as seen in

Figure 5, S4 used representations for two different variables (mean and maximum PM10 levels) to answer Q1-'Level of danger'. The student noted the variability in the data from the bar chart showing the maximum PM10 levels by location and year by referencing to 'noise' (i.e., "anomalies") in the raw data. When he represented the distribution of mean PM10 levels by location and year in a bar chart, the PM10 levels were found safe. On the other hand, in Q2-'Rising overtime', all students who used either some measures or visualization of data focused on the consistency in increase of mean and/or median values over three years to identify the trend in PM10 levels over time for each location. For instance, S6 (Figure 5) used a table of statistical measures to note the consistency in increase of the mean and median PM10 levels over time and considered no variation in the rise of PM10 levels from one year to next. Even though students calculated the measures of variability, such as standard deviation and interquartile range, the variability around the mean or median within a year was not expressed in their claims.



S4.Q1-'Level of danger':"From the second graph [*on the right*], one could claim that in RAMM (2016) PM10 levels were at dangerous levels and thus yes, PM10 levels have reached dangerous levels. However, it is possible that these occurrences are inaccurate; By looking through the raw data, all high occurrences swiftly return to their lower level recordings. Nonetheless, the occurrences are rare and should be considered anomalies. However, by using the mean averages of all recorded values, show that all locations and years are considered "Low" and thus are not dangerous. With the highest mean value being Alphington (2017) with only 20.5 it's safe to assume based on the provided data, that the PM10 levels are safe."

	RAMM 2015	RAMM 2016	RAMM 2017	Alphington 2015	Alphington 2016	Alphington 2017
Mean	15.4	17	19.4	15.2	17.1	20.5
Median	13.5	15.1	16.8	13.6	15.5	17.8
Standard Deviation	9	10.7	10.6	8.6	9.7	11.6
IQR	12.179	12.19125	13.967	11.186751	12.3215	14.8645

S6.Q2-'Rising overtime': "Looking at the means and medians for the two sites over the three years we can see there is definitely a gradual increase in PM10 levels. The RAMM site goes steadily up 4 points and at Alphington it rises steadily 5 points. The consistent increase year on year reduces the possibility of it being a random increase over two years and in-fact a trend which will increase next year and will have likely been increasing in prior years up to 2015. More data will need to be collected in coming years to help prove this trend."

Figure 5. Sample student responses (S4 and S6) for Q1-'Level of danger' and Q2-'Rising overtime'

Unlike in Q1-'Level of danger' and Q2-'Rising overtime', seven students were inclined to use various data representations, such as table, histogram, bar chart, dot plot, scatterplot, pie chart, time series plot, and line graph, when answering the other two questions, 'What time of day are PM10 levels the highest?'[Q3-'Time of day'] and 'Is there any correlation between PM10 levels at the two sites?'[Q4-'Correlation']. Q3-Time of Day requires students to consider hourly PM10 levels data in aggregate by the categories of the time of the day (morning, afternoon, evening, and night) given in the dataset. It was the one with the most varying answers, depending on how students structured the data to explore the variability in the data set.

For example, S1 used both tables showing different sets of raw data for different locations and times and histograms for two locations in the same year ~~from which he gained different insights into~~ to show ? what time of day PM10 levels were at their highest (Figure 6). He noted that there was no obvious answer to this question due to varying patterns seen in the visualizations. Looking at the raw data also allowed him to acknowledge some anomalies in the PM10 values. These extreme values were evaluated as very rare occasions compared to the rest of the data. The histograms helped the student

interpret the range of values where most of the data were clustered to compare the PM10 levels in both locations. This way the student was able to account for variability around the center when interpreting the distributions of data.

S1.Q3-'Time of day': "There is no obvious answer to this, the results are relatively spread out. However, there are multiple occasions where there are high PM10 levels during the morning and afternoon, but these are not overly high and are nothing to worry about. There were two instances of more extreme PM10 levels, one in RAMM 2016 and the other in Alphington 2017. Both are in the month of November, the data is shown [*on the right*].

Date/Time	RAMM 2016
29th Nov. 17:00	50.473
29th Nov. 18:00	63.67
29th Nov. 19:00	79.536
29th Nov. 20:00	57.114

There are two occasions in the evening where there is a high level of PM10.

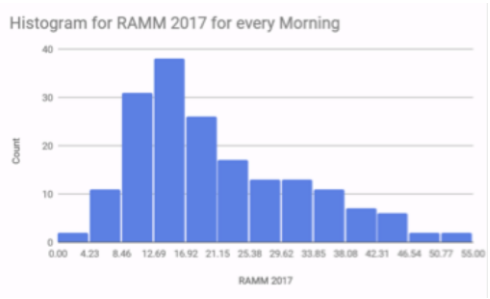
Date/Time	Alphington 2017
3rd Nov. 18:00	40.254
4th Nov. 19:00	55.199
5th Nov. 20:00	78.776
6th Nov. 21:00	58.855
6th Nov. 22:00	57.358

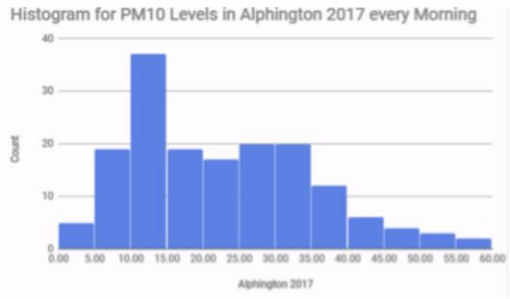
On the same month the next year, there was similar PM10 levels around the same time, but at a different location.

These are well above average values for the PM10 levels but it was only for one evening/night of two separate years. For the rest of the year, the only time values were high were in the morning. afternoon and maybe in the evening (people travelling from and to work) as shown [*on the right*].

Date/Time	RAMM 2016
29th Nov. 7:00	27.102
29th Nov. 8:00	30.107
29th Nov. 9:00	33.44
29th Nov. 10:00	32.812
29th Nov. 11:00	41.977
29th Nov. 12:00	36.92

During the morning hours the PM10 levels spike every so slightly above average, maybe this is due to the amount of traffic in that area from everyone going to work.



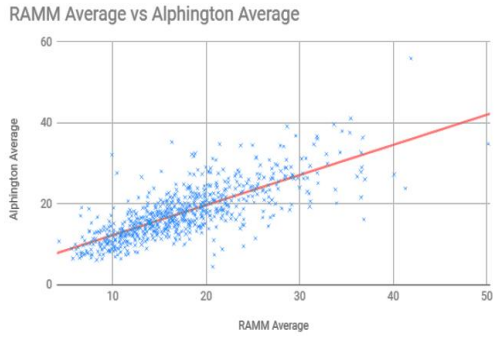


These histograms tell us that the most common values for the mornings of RAMM in 2017 are between 8.46 and 21.15 whereas in Alphington 2017 they are more distributed between 5 and 35, which interestingly is a much higher range meaning it is much more consistent to have higher levels of PM10 in Alphington."

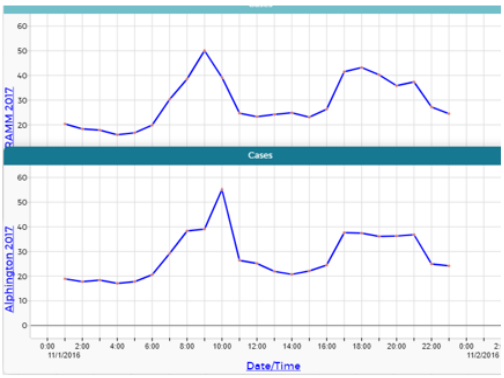
Figure 6. Sample student response (S1) for Q3-'Time of day'

Unlike in Q3-'Time of day', responses to Q4-'Correlation' ('Is there any correlation between PM10 levels at the two sites?') were all consistent based on choices of different visual displays to look at the trends in the mean PM10 values over three years for each location (with the exception of S10 using raw data in 2017 for each location). As seen in these examples below (Figure 7), data visualizations, such as scatterplots and line

graphs, displaying either average values (**S1.Q4-'Correlation'**) or raw data (**S10.Q4-'Correlation'**) for PM10 levels in each location helped students to see a similar trend or pattern in the data. While S1 focused on the direction of the trend line in the visualization (scatterplot and line of best fit), S10's data displays (line graphs) allowed him to compare the peaks and drops in the same time intervals over 24 hours for each location from November 1st in 2017. The other types of visualizations used by other five students, including combined time series, clustered bar graph, dot plot, and pie chart, displayed average PM10 levels for each location over time and allowed them to compare the patterns, i.e., the change in the average values over three years, in the data in both sites. It appeared that all students but three chose to visualize the data to explore if there was any relationship between PM10 levels at both locations in the city. While the scatterplot helped a student to determine the positive relationship expressed by the line of best fit between the two variables, the other data representations enabled students to see if the patterns for each location coincided over time.



S1.Q4-'Correlation': "I've used the averages for each of the days in November for both RAMM and Alphington for all years to get a rough idea if there is any correlation between the PM10 levels at the two sites. From the results shown on the graph I can see there is a clear positive correlation and both RAMM and Alphington follow the same pattern, which explains the trend line going in an upwards diagonal direction."



S10.Q4-'Correlation': "There is correlation between the two sites. For example, comparing the data points from the 1st November 2017 at both sites we can see they both have peaks around 9-10am and 5-6pm followed by drops in PM10 levels. This is most likely due to rush hour, where there are more vehicles on the street contributing to PM10 levels."

Figure 7. Sample student responses (S1 and S10) for Q4-'Correlation'

As seen in the examples presented above across four questions investigated in the 'explore data' phase of the DA cycle, students chose to use either statistical measures (mainly averages) or various data visualizations to interpret PM10 level data by considering variation both within and between each location and year. While summary tables including several statistical measures for center and variation seemed to allow students to attend to patterns over time in answering Q1-'Level of danger' and Q2-'Rising overtime', Q3-'Time of day' and Q4-'Correlation' required students to move beyond these summary statistics by visualizing the data with different graphs. For example, in Q3-'Time of day' students needed to break down the PM10 level data (either mean values or raw data) from both locations over three years by another data category, 'time of the day', to see what time of day PM10 levels tended to be the highest. Similarly, Q4-'Correlation' mostly led them to either consider distribution of values (raw data or means) over a day or a month to identify a pattern between each location or attend to the variation in average PM10 levels across years.

Students' expressions of (un)certainty and use of context in making inferences

In addressing students' informal inferences on their reports as part of the 'draw conclusions' phase of the DA cycle, we focus on two emerging themes from our analyses: students' certainty/uncertainty in their conclusions and their use of contextual knowledge.

Students' certainty in making their conclusions was salient primarily in Q2-'Rising overtime' and Q4-'Correlation' responses and somewhat in Q1-'Level of danger' responses. In Q1-'Level of danger', it was mainly in relation to noticing rare anomalies or peaks in the data when five of the students also looked at the dataset for extreme PM10 values. As seen in the example below, S7's confidence in his claim was not influenced by these rare occasions (i.e., the peak PM10 levels for both locations on a particular evening in 2017) noted in the data:

> ...Despite the mean average sitting in the low band for both sites, the peak PM10 value for Alphington reached 60.444 in 2017, and for RAMM it reached 78.776 on the same evening. These levels reach well into dangerous levels of Moderate and High bands. However, the levels rarely reached close to this throughout the 3 years, this is most likely because some kind of event was happening in the city on 03/11/17. (**S7.Q1-'Level of danger'**)

Here, noticing that particular anomaly (the two PM10 values considered at dangerous levels) in the dataset seemed to conflict with S7's earlier claim about the mean PM10 values being in the safe levels. However, the student tried to justify that these were only rare occasions in the whole dataset by referring to some causal event linked with the context.

In Q2-'Rising overtime' and Q4-'Correlation' particularly, students' confidence in their conclusions tended to include some qualifiers for their strong certainty in reference to trends in the data, e.g., "*can clearly see*", "*quite certainly*", "*definitely*", "*there is no argument*", "*this graph clearly shows*" and "*there is an undeniable positive*

correlation". All students but one (S3) were certain or highly certain about their informal inferences (e.g., **S4.Q4-'Correlation'**: "From the data and the two charts, both locations do appear to be similar in results. Due to this, I would say there is a *fairly strong correlation* between them. This is likely due to their locations being close to one another."). However, in answering Q2-'Rising overtime', S3 seemed to be not completely certain in his conclusion citing the limitation of data:

> The graph shows that for both sites, the mean PM10 is higher after each successive year, indicating that PM10 levels in these areas are indeed rising over time. This is however only including the months of November as we only have data for that month in particular, so each successive November has higher PM10 values on average, but whether PM10 levels are rising overall per year is *not a fully assured statement* to make unless we assume the rest of the year's results are alike. (**S3.Q2-'Rising overtime'**)

As seen in S3's response, uncertainty in the inference was expressed with reference to the dataset including the records only in a particular month across three years. Awareness of data limitation in relation to small sample size was also evident in three students' responses in Q2-'Rising overtime', but they did not use a qualifier to express an uncertainty in their inferential statements, such as "Despite having a very small dataset this shows that PM10 levels are on the rise." (**S10.Q2-'Rising overtime'**).

Students' use of context generally appeared in Q3-'Time of day' responses in which students gave very different answers to what time of the day PM10 levels were at their highest. Perhaps due to the lack of a clear pattern in the data, students tended to bring in their contextual knowledge to provide a justification for their informal inferences, as seen in the following example:

> PM10 levels are at their highest for both sites during the Evening. As mentioned before, PM10 levels reached 60.444 in Alphington and 78.776 at RAMM on 03/11/17. Both of these recordings were taken during the evening at 8pm. This could be explained through specific weather patterns such as pressure levels causing particulates to be trapped in the city. It could also be explained due to popular events taking place in the city mixed with the complete coincidence that more people were travelling by diesel vehicle that day than usual. (**S7.Q3-'Time of day'**)

Moreover, students claimed different times of the day for the highest level of PM10 with the knowledge of context used to justify their inferences and as a story behind the data. For instance, S7 argued that evening was the highest time of the day, as seen in the above quote (**S7.Q3-'Time of day'**), whereas S8 decided that it was the mornings with the highest PM10 level, supported by the contextual knowledge:

> PM10 levels are, on average, frequently highest in the morning, as shown by the table below. Despite some values being higher (e.g. "Night, 23.8" being higher than "Morning, 16.4"), the highest levels in ½ (3/6) of the months were recorded in the morning (Green Cells). This is likely due to the morning commute, where not only are more vehicles using the roads (Producing more pollution) but these cars are likely to have started cold (Meaning the catalytic converter in the exhaust is unable to reduce the amount of particulate matter released as effectively until the car warms up again). (**S8.Q3-'Time of day'**)

There were other occasions that the use of contextual knowledge allowed students to go beyond the data to make predictions, such as the response of **S7.Q3-'Time of day'** above, and the knowledge of context helped to explain the limitation of a claim as seen in another response:

> However, there are small places where the data does not correlate, where we see a rather flat afternoon peak at Alphington compared to a higher spike at RAMM. This

> could be due to location difference, where RAMM is much closer to the central hub of the city. (**S10.Q4-'Correlation'**)

Thus, as seen in the examples above, how students utilized their knowledge about context in their inferences varied across the four questions. The use of context became more apparent, particularly when there was no clear pattern in the data and students needed to support their claims.

Discussion and conclusion

In this article, we examined how students explored data and drew conclusions from a large multivariate dataset provided by third parties with the use of technology tools, as part of the DA cycle during an activity designed to stimulate competencies like statistical literacy, ICT literacy, critical thinking, and collaboration and communication, based on the DA framework (Figure 1). From a data modeling perspective, we were particularly interested in investigating (1) students' ways of measuring and visualizing to identify and interpret trends/patterns in the data and (2) their approaches to drawing inferences. The problem presented to the students involved a real dataset comprised of 720 cases with hourly PM10 data from November 2015, 2016 and 2017 and time of the day (morning, afternoon, evening, night) in two different locations in the city. The inherent variation, missing values, anomalies, and lack of clear trends in the sample data required students to interpret variability and acknowledge uncertainty in making inferences.

The use of statistical measures to summarize distribution properties, such as center and variability, and data representations to visualize those properties are found to be useful models for comparing and reasoning about data distributions in the existing literature (Büscher & Schnell, 2017; Doerr et al., 2017; English & Watson, 2018; Fielding-Wells, 2018). Within the 'Explore data' phase of the DA cycle in this study,

students primarily relied on statistical summaries (particularly averages) to see a trend in the data when exploring whether PM10 levels reached a dangerous level (Q1-'Level of danger') and have been rising over time (Q2-'Rising overtime'). Once students determined the safe level of PM10, considering the mean PM10 values by year and location seemed to make sense to them. Hence, the inferences required by these questions facilitated students' use of average measures in meaningful contexts as seen in other studies aimed at promoting students' informal inferential reasoning (Makar, 2014) and modeling (Büscher & Schnell, 2017; English & Watson, 2018) to develop their understandings of statistical concepts. Some of the students also looked at the individual recordings of PM10 levels in the dataset and noted a few anomalies at dangerous levels. They commented on those anomalies while explaining how average recordings of PM10 values were considered in response to Q1-'Level of danger' (see example **S4.Q1-'Level of danger'** in Figure 5). On the other hand, almost all students computed measures of variability, such as standard deviation and interquartile range, but they did not actually use them as part of their analysis in their reports. This seems consistent with other reports stating that typically students learn how to compute measures of variability, but they have difficulty in understanding what these measures represent and using them in connection with other statistical concepts when interpreting the data (Garfield & Ben-Zvi, 2008).

In contrast to student responses in Q1-'Level of danger' and Q2-'Rising overtime', there was a tendency to use data visualizations for looking at patterns or trends in data when students explored what time of the day PM10 levels were at their highest (Q3-'Time of day') or if there was a correlation between PM10 levels at two locations (Q4-'Correlation'). As argued by Ridgway et al. (2017), the use of data visualizations facilitated students' exploratory analysis of data to identify patterns or trends in the given data. More specifically, various data representations, such as histogram, dot plot, line

graph, scatterplot and so on, used by the students in answering these two questions allowed them to visualize variability and identify patterns in the data (Doerr et al., 2017; Fielding-Wells, 2018). It was through these visualizations that students tended to account for variability. For example, in Q3-'Time of day', S1 referred to the range of the cluster in the middle when comparing histograms of the PM10 values in the morning for each location. Without measuring variability, the student could provide a plausible explanation for why it was more likely to have higher PM10 levels in the morning in Alphington than in RAMM. Also, many students observed a strong correlation between PM10 levels at two locations by looking at the patterns in the data that more or less coincided over time in their graphs for each location. As argued by Makar and Rubin (2018), this was evidence that students allowed for variability when interpreting data.

During the 'Draw conclusions' phase of the DA cycle, in which students made inferences from the provided data, they tended to express their (un)certainty in their conclusions using probabilistic language or qualifiers for strong confidence especially in Q2-'Rising overtime' and Q4-Correlation. Acknowledgment and articulation of uncertainty are expected in making inferences based on a sample (Ben-Zvi et al, 2011; English & Watson, 2018). In this study, all students except one drew their conclusions with some or high certainty, which was evident in their use of qualifiers for strong confidence (e.g., **S4.Q4-'Correlation'**) in responding to Q2-'Rising overtime' and Q4-'Correlation'. Although three of the students noted data limitation due to small sample size in their statements, they did not articulate any uncertainty in drawing their conclusions based on this dataset. Only one student (S3) acknowledged the uncertainty in his inference by noting that the dataset only included November data for each year.

Furthermore, we found that students' use of contextual knowledge became quite apparent in their informal inferences when responding to Q3-'Time of day' in particular.

One reason for that is probably because the data did not provide a clear pattern to determine what time of the data PM10 levels were at their highest. So, the students relied on their contextual knowledge to justify their inferences based on their statistical analysis. In line with findings of Langrall et al. (2011) and Pfannkuch (2011), a consideration of context appeared to tell a story behind the data, go beyond the data to make predictions, and explain the limitation of a claim in student responses in Q3-'Time of day'.

Prior research on students' data modeling has been conducted with young students using relatively small datasets. The present study extended this work through engaging 17-18-year-old students, who have not studied advanced statistical topics, in the data modeling process with the aim to develop their data analytics skills. Moreover, by encouraging students to explore large multivariate datasets supplied by third parties using technology tools, we focused on how students identified possible relationships between the dataset variables to draw inferences in the context of air pollution to connect data, chance and context. As argued by English and Watson (2018), data modeling problems that are complex and ill-structured foster students to make sense of problem context, to structure and visualize data, and to draw inferences from data. In this study, we used a sample from a real dataset containing missing values and anomalies to challenge students to consider data collection and the reliability and accuracy of data supplied by third parties when making decisions with acknowledgement of uncertainty. However, the dataset was presented to students with a variable specifying what time of day (morning, afternoon evening) a PM10 level was measured based on the time collected originally. Instead, as seen in exploration of large multivariate datasets (Erickson, Finzer, Reichsman, & Wilkerson, 2018), students could be encouraged to structure the given data on their own as they see the need and decide how they would group the data. This way of creating a new variable from existing data might emerge during the data modeling process when the

opportunity is given (see English & Watson, 2018) and would be a necessary skill to develop as part of the second phase of the DA cycle 'Consider data'.

Overall, this study suggests that implementing data modeling tasks coupled with skills and competencies described in the DA framework can start to promote students' informal statistical inferences with considerations of variability, context, and uncertainty. It can be expected that accounting for variability does not naturally occur to students, even when they are able to compute measures of variability in their analysis. However, encouraging students to visualize distributions of data can help them consider variability in their exploration of a multivariate dataset. When engaging students in sample-to-population inference, another type of variability that can emerge is sampling variability, which was not, however, addressed in students' responses in this study. Although some students noted the limitation in their statements due to the small dataset (including November data only), sampling variability was not considered in making inferences. It can be recommended that in order to see the role of sampling variability, students should be given an opportunity to sample from the same population and explore different samples to draw inferences. Acknowledging uncertainty is one of the key elements of informal statistical inference. Given the students' difficulties in expressing uncertainty in their inferential statements based on the given sample in the current study, exploring multiple samples from the same population for making informal inferences can be suggested for initial data investigations.

There are some limitations to this research. First, the study was conducted with students who were studying vocational subjects, such as Informational Technologies for professional qualifications. Students from other types of upper secondary schools would have enhanced the utility of findings. Second, the duration of the lesson was limited to reviewing the statistical content and completing the task through the phases of the DA

cycle. Students would have benefited from applying their knowledge and skills gained from the described task to new contexts. However, this study was the first iteration of the pilot project by the teacher aimed to develop students' DA skills and competencies. The findings point out the potential of engaging students in a data modeling task that involves exploring a large multivariate dataset supplied by third-party providers with the use of technological tools. This data modeling approach appears to provide students opportunities with developing skills and competencies related to DA, especially in exploring data, drawing conclusions, using statistical concepts and ICT tools, critical thinking and communication of results.

The study has three major implications for researchers and teachers seeking to design activities to enhance students' DA skills to solve a real-world problem using a large multivariate dataset supplied by third-party providers. Firstly, more explicit and extended discussion of how data are sampled is needed for consideration of variability, in particular sampling variability, and expression of uncertainty in making sample-to-population inferences. Secondly, when students are introduced to the DA cycle to explore a large multivariate dataset, they are likely to need more guidance to focus their investigations with semi-structured questions leading them to interpret variability and acknowledge uncertainty in making inferences. Thirdly, encouraging students not only to compute statistical measures but also to visualize distributions can facilitate accounting for variability in drawing inferences. This study focuses on students' explorations of a sample of data provided to them. A potential venue for further research is to investigate how students engage in DA cycle through data modeling when they can select multiple samples to make inferences and discuss the conclusions.

Acknowledgement

[…]

References

Arnold, P. (2013). *Statistical investigative questions. An enquiry into posing and answering investigative questions from existing data*. Unpublished doctoral thesis. The University of Auckland.

Bakker, A., & Derry, J. (2011). Lessons from inferentialism for statistics education. *Mathematical Thinking and Learning, 13*(1-2), 5-26.

Ben-Zvi, D. & Garfield, J. (2004). Statistical literacy, reasoning, and thinking: Goals, definitions, and challenges. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 3–25). Dordrecht, The Netherlands: Kluwer.

Ben-Zvi, D., Aridor, K., Makar, K., & Bakker, A. (2012). Students' emergent articulations of uncertainty while making informal statistical inferences. *ZDM, 44*(7), 913-925.

Büscher, C., & Schnell, S. (2017). Students' emergent modelling of statistical measures–a case study. *Statistics Education Research Journal*, *16*(2), 144-162.

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Shauble, L. (2003). Design experiments in educational research. *Educational Researcher, 32*(1), 9-13.

Common Core State Standards Initiative (2010). *Common core state standards for mathematics*. Washington, DC: National Governors Association for Best Practices and the Council of Chief State School Officers. Retrieved from http://www.corestandards.org/assets/CCSSI_Math%20Standards.pdf

Department for Education (2014). The national curriculum in England: Key stages 3 and 4 framework document. Retrieved from https://www.gov.uk/government/publications/national-curriculum-in-england-secondary-curriculum

Doerr, H. M., delMas, R., & Makar, K. (2017). A modeling approach to the development of students' informal inferential reasoning. *Statistics Education Research Journal*, *16*(2), 86-115.

English, L. D. (2010). Young children's early modelling with data. *Mathematics Education Research Journal*, *22*(2), 24-47.

English, L. D., & Watson, J. (2018). Modelling with authentic data in sixth grade. *ZDM*, *50*(1-2), 103-115.

Erickson, T., Finzer, B., Reichsman, F., & Wilkerson, M. (2018). Data Moves: one key to data science at school level. In M. A. Sorto, A. White & L. Guyot (Eds.), *Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10)*. Retrieved from http://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_9B3.pdf?1531454621

Fielding-Wells, J. (2018). Dot plots and hat plots: Supporting young students emerging understandings of distribution, center and variability through modeling. *ZDM*, *50*(7), 1125-1138.

Franklin, C., Kader, G., Mewborn, D. S., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2005). *Guidelines for assessment and instruction in statistics education (GAISE) report: A curriculum framework for K-12 statistics education*. Alexandria, VA: American Statistical Association.

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. International statistical review, 70(1), 1-25.

Garfield, J., delMas, R., & Chance, B. (2003). Web-based assessment resource tools for improving statistical thinking. *Paper presented at the annual meeting of the American Educational Research Association*, Chicago.

Garfield, J. & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: connecting research and teaching practice*. New York, NY: Springer.

Kazak, S., Pratt, D., & Gökce, R. (2018). Sixth grade students' emerging practices of data modeling. *ZDM Mathematics Education, 50*(7), 1151–1163.

Langrall, C., Nisbet, S., Mooney, E., & Jansem, S. (2011). The role of context expertise when comparing data. *Mathematical Thinking and Learning, 13*(1-2), 47-67.

Lehrer, R., & Schauble, L. (2004). Modeling variation through distribution. *American Education Research Journal, 41*(3), 635-679.

Lehrer, R. & English, L. (2018). Introducing children to modelling variability. In D. Ben-Zvi, K. Makar and J. Garfield (Eds.), *International Handbook of Research in Statistics Education*, (pp. 229-260). NY: Springer International Publishing.

Lesh, R. A., & Doerr, H. M. (Eds.). (2003). *Beyond constructivism: Models and modeling perspectives on mathematics problem solving, learning, and teaching*. Mahwah, NJ: Lawrence Erlbaum Associates.

Makar, K. & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal, 8*(1), 82–105. [Online: http://iase-web.org/documents/SERJ/SERJ8(1)_Makar_Rubin.pdf]

Makar, K. (2014). Young children's explorations of average through informal inferential reasoning. *Educational Studies in Mathematics*, *86*(1), 61-78.

Makar, K., & Rubin, A. (2018). Learning about statistical inference. In D. Ben-Zvi, K. Makar and J. Garfield (Eds.) *International Handbook of Research in Statistics Education* (pp. 261-294). NY: Springer International Publishing.

Parlett, M., & Hamilton, D. (1972). *Evaluation as illumination: A new approach to the study of innovatory programs, Occasional Paper no 9*. University of Edinburgh,

Centre for Research in the Educational Sciences, Edinburgh. (Retrieved from ERIC database (ED167634)).

Pfannkuch, M. (2011). The role of context in developing informal statistical inferential reasoning: A classroom study. *Mathematical Thinking and Learning, 13*(1-2), 27-46.

Pfannkuch, M., Ben-Zvi, D., & Budgett, S. (2018). Innovations in statistical modeling to connect data, chance and context. *ZDM, 50(7), 1113-1123.*

Picciano, A. G. (2012). The evolution of big data and learning analytics in American higher education. *Journal of Asynchronous Learning Networks, 16*(3), 9-20.

Ridgway, J. (2016). Implications of the data revolution for statistics education. *International Statistical Review, 84*(3), 528-549.

Ridgway, J., Nicholson, J., Campos, P., & Teixeira, S. (2017). *Tools for visualising data: A review*. IASE Satellite Conference.

Royal Society (2016). *Data analytics: the skills need in STEM*. Conference report organised in partnership with the Royal Statistical Society held on 16 November 2016.

Shinohara, M., & Lehrer, R. (2017). Narrating lines of practice: Students' views of their participation in statistical practices. In *Proceedings of the 10th Statistical Reasoning, Thinking and Literacy Forum, Rotorua, New Zealand* (pp. 15–33). Auckland: SRTL.

Watson, J., & English, L. (2018). Eye color and the practice of statistics in Grade 6: Comparing two groups. *Journal of Mathematical Behavior, 49*, 35-60.

Wild, C. J. (2018). *The place of data analysis in problem solving*. Unpublished manuscript. The University of Auckland, New Zealand.

Wild, C. J. & Pfannkuch, M. (1999). Statistical thinking in empirical inquiry. *International Statistical Review, 67*(3), 223-248.

Wild, C. J., Utts, J. M., & Horton, N. J. (2011). What is statistics? In D. Ben-Zvi, K. Makar and J. Garfield (Eds.) *International Handbook of Research in Statistics Education* (pp. 5-36). Springer International Handbooks of Education. DOI 10.1007/978-3-319-66195-7