Published in final edited form as:

Int J Math Educ Sci Technol. 2011 September 1; 42(6): 789–829. doi:10.1080/0020739X.2011.562315.

Web-based tools for modelling and analysis of multivariate data: California ozone pollution activity

Ivo D. Dinov^{a,b,*} and Nicolas Christou^a

^aStatistics Online Computational Resource, University of California, 8125 Mathematical Science Building, Los Angeles, CA 90095, USA

^bCenter for Computational Biology, University of California, 8125 Mathematical Science Building, Los Angeles, CA 90095, USA

Abstract

This article presents a hands-on web-based activity motivated by the relation between human health and ozone pollution in California. This case study is based on multivariate data collected monthly at 20 locations in California between 1980 and 2006. Several strategies and tools for data interrogation and exploratory data analysis, model fitting and statistical inference on these data are presented. All components of this case study (data, tools, activity) are freely available online at: http://wiki.stat.ucla.edu/socr/index.php/SOCR_MotionCharts_CAOzoneData. Several types of exploratory (motion charts, box-and-whisker plots, spider charts) and quantitative (inference, regression, analysis of variance (ANOVA)) data analyses tools are demonstrated. Two specific human health related questions (temporal and geographic effects of ozone pollution) are discussed as motivational challenges.

Keywords

statistics education; science education; data analysis; case study; applets; ozone pollution; SOCR

1. Introduction

1.1. Motivation and background

The use of simulated, observational or research-derived data has a significant effect on improving science literacy and enhancing the learning experiences in multidisciplinary education at different levels [1-5]. A number of factors may be inhibiting the integration of real datasets in the modern science education curricula. Examples of these limitations include the availability of interesting motivational datasets, access to appropriate tools for data manipulations and technical skills for interrogating the data to address specific driving scientific questions. This article addresses these challenges by posing a pertinent health-related question, providing access to appropriate data and illustrating the use of open-access tools for data analyses and scientific inference.

Ozone pollution is associated with many human health problems [6]. For instance, it is correlated with the prevalence of chronic lower respiratory tract symptoms in children [7], associated with new onset asthma in adolescents [8,9] and implicated in development of adult-onset asthma in non-smoking adults [10].

This article demonstrates an end-to-end hands-on example of a multidisciplinary scientific problem followed by description of data acquisition, modelling, interrogation, analysis and scientific inference using ozone pollution as a driving biological challenge. The following investigative processing protocol (pipeline) is discussed below:

Problem → Data → Model → Analysis → Inference/Decision-making

Some of the specific aims of this case study are to describe the processes of:

- Formulation of meaningful research questions or hypotheses;
- Identification and collection of appropriate datasets;
- Use of appropriate technology for data manipulation and interrogation, as well as for addressing the driving research hypotheses;
- Scientific inference.

1.2. Formulation of meaningful research questions and hypotheses

Scientific inquiries are typically motivated by observational experiences (inductive approach) or by abstract investigations (deductive approach) of physical, chemical, biomedical, clinical, environmental or other processes or natural phenomena. These two approaches are complementary, provide valuable means for new discoveries and their validation, and are employed synergistically in science education for motivation and improvement of learning experiences. This article demonstrates a common observational scientific research protocol in the context of science education using the inductive approach.

The general motivational questions this study concentrates on are the temporal and geographic effects on ozone pollution in California. Ozone is both beneficial and harmful to life as it forms in the stratosphere and serves as a vital and effective protective barrier from the sun's ultraviolet rays. At high altitudes, depletion of the ozone layer, mainly caused by the emission of various chemical pollutants (e.g. chlorofluorocarbons), is associated with an increase of cancer rates due to single nucleotide polymorphism mutations and failures in deoxyribonucleic acid (DNA) repair [11]. However at lower altitudes, high concentrations of ozone (ground-level ozone pollution) may be harmful to human health. Ozone pollution is a byproduct of the interaction between sunlight, hydrocarbons and nitrogen oxide (emitted by burning fossil fuels). During high temperature and low wind conditions low altitude ozone levels can reach unhealthy levels. The US Environmental Protection Agency's Air Quality Index (AQI) is a marker of severity of the ozone pollution. AQI provides both a qualitative (colour) and a quantitative (particles per million, *ppm*) measures, Table 1.

In 2010, the Obama administration proposed to set the standard for ground-level ozone of no more than 0.060 to 0.070 ppm, which is expected to be phased in over the next two decades. Previously, in 2005, the Bush administration imposed an upper limit of 0.075 ppm, and the Clinton administration introduced an upper limit of 0.084 ppm in 1997. The monotonic downward trend in these ozone pollution guidelines is indicative of the increased scientific knowledge and our awareness about the health related affects of ozone pollution over the last two decades.

There are many health-related hypotheses that can be formulated and addressed in this setting. Examples of two specific questions that may be used for problem motivation include:

1. Are there temporal changes in California Ozone pollution?
2. What is the geographic distribution of the California Ozone pollution and is it changing with time?

1.3. Identification and collection of appropriate datasets

Once we have a concrete motivational challenge we need to identify an appropriate dataset that will help us address the problem. There are many ways to obtain such data – experimental designs, observational and longitudinal studies, simulations, etc. In the case of ozone pollution, the Environmental Protection Agency (www.EPA.gov) and the California Air Resources Board (CARB, www.arb.ca.gov) provide such data. In this activity we used data from CARB (http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_121608_OzoneData).

This dataset includes the ozone pollution measurements (in ppm) for 20 geographic locations within the State of California for the period between 1980 and 2006. This is a complex dataset where we chose to focus only on one interesting measurement variable OZMAX1HR (Maximum 1-hr Ozone average concentration). The complete collection of the data also contains other measurement variables like: OZ4HI3YR (Ozone 3 year 4th high daily maximum over 1 h, in ppm); OZXH1HST (count of hours Ozone exceeds the State 1 hour standard (0.095 ppm)); ZXH1HNA (count of hours Ozone exceeds US national 1 h standard (0.125 ppm)); and OZEX1HST (number of days Ozone count exceeds State 1 h standard (0.095 ppm)).

Several meta-data variables were also available:

- LOCATION = geographic location, Figure 1. An interactive geo-map of these locations is available here: www.socr.ucla.edu/docs/resources/SOCR_Data/SOCR_OzoneData_GoogleMap.html
- YEAR = measurement year
- Measurement month: M1: MTH_1 (January), M2: MTH_2 (February), M3: MTH_3 (March), M4: MTH_4 (April), M5: MTH_5 (May), M6: MTH_6 (June), M7: MTH_7 (July), M8: MTH_8 (August), M9: MTH_9 (September), M10: MTH_10 (October), M11: MTH_11 (November), M12: MTH_12 (December)
- ANNUAL = annual average of the ozone measurement for that location/year
- COMP_DAYS = Number of Days with Complete Data
- HI_COVER = High Coverage (coverage during typical periods of high concentration)
- COMP_SITES = Number of Sites with Complete Data
- LATITUDE = location latitude
- LONGITUDE = location longitude
- ELEVATION (Elev) = location altitude (elevation), height above sea level in meters.

The entire dataset is represented as a spreadsheet table of size 540(rows) $\times$ 23(columns), see URL above. The rows represent individual observations (at 20 locations over a period of 27 years), and the columns contain the quantitative data for the variables described above.

1.4. Use of appropriate technology

For the purposes of this case study we employ the web-based tools developed by the UCLA Statistics Online computational Resource (www.SOCR.ucla.edu). However, many other tools for data representation, analysis and visualization would provide similar functionalities. We have chosen the statistics online computational resource (SOCR) Charts [12,13] and SOCR Analyses [14] because they facilitate open anonymous access and can be

directly integrated in, or embedded with, other web-based activities, computational resources, visualization tools and web-services.

1.5. Inference and decision-making

Scientific inference is the final stride in the general *Problem* $\rightarrow$ *Inference* pipeline. The entire data analysis process typically includes all data-driven procedures and the inference arising from the intrinsic or extrinsic random variation present in the (sample) data. The goal of the final inference step is to provide reasonable answers to the initial driving motivational problems and give clues for what may be expected as outcome in other similar situations. An important aspect of this process is the fact that the final scientific inference has population-wide relevance; however the inference is derived based on sample data (representative of the specific population of interest). In the ozone pollution case, and the two specific driving motivational questions, our inference may include providing sufficient (deductive) information for taking appropriate action based on the *temporal* or *geographic* trends of the ozone pollution in California.

2. Data and case study design

Table 2 shows a small fragment of the data and the complete ozone pollution dataset is freely available online at http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_121608_OzoneData

To address our two health-related questions (temporal changes and geographic distribution of the California Ozone pollution) we will employ exploratory data analyses (EDAs) (box-and-whisker charts, motion-charts, spider charts) as well as quantitative analyses (inference, regression, analysis of variance (ANOVA)). Figure 2 illustrates the design of this ozone pollution case study referring to the two motivational problems, appropriate SOCR exploratory and quantitative data analysis tools, suggested inference outcomes and applications of this study. Indeed, these steps represent examples of research questions that may be formulated, instances of tools that can be employed to solve the fundamental challenges and suggested inference-based decision-making applications. The reader is encouraged to think of other appropriate data, questions, tools and inference.

3. Pedagogical use

Data-driven interactive demonstrations of various basic or applied scientific concepts in the classroom serve several complementary purposes. Such demonstrations ground the newly presented materials, techniques or concepts as practically useful, attract students' attention and increase learning motivation [1]. In addition, data-driven and technology-enhanced illustrations facilitate holistic conceptual understanding, bridge between different disciplines, promote extracurricular knowledge and encourage students to experiment and empirically validate various alternative scenarios (experiment confirmatory experiences or construction of counterexamples).

The analysis examples shown below were chosen because they are frequently useful in practice and can easily be replicated by students during learning and preparation, and by instructors in the classroom. Teachers may choose to extend, modify or supplement these examples in various ways to fit their specific course syllabi, curricular settings, student maturity levels or learning styles. For instance, both the selection of a driving motivational problem and the data collection strategy could be excellent sources to stimulate students' creativity and promote their individual interests. Many interesting datasets are available online (e.g. http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data, <http://TheInfo.org>).

3.1. Exploratory data analyses (EDA)

Many graphical displays of the data improve the intuitive understanding of the data characteristics and provide qualitative description of the overall dataset. All EDA strategies attempt to visually summarize the random variation present in the dataset and expose intricate trends and relations in the data. EDA discussions engage students in conceptual and methodological topics and provide some quick cues to the sample and population characteristics of interest. We demonstrate three such EDA tools using SOCR Charts, http://socr.ucla.edu/htmls/SOCR_Charts.html, [12,13].

Spider Charts enable the display of several different variables related to one item as a web or radar plot with multiple axes corresponding to different factors in the data. In spider charts, a tabular numerical value is represented as the distance between the chart centre and a point on any of the spider-web axes which represent different factors in the data. Low and high data values within a factor generate points closer or farther from the centre of the chart, respectfully. Figure 3 shows a spider chart of ozone pollution data across years, factors, (angularly) for all 20 locations (radially). A clear pattern emerges of the differences between the ozone pollution levels in the 1980's and their counterparts in the 2000's.

Box-and-whisker charts are standard data graphing plots representing a convenient way of graphically depicting outliers and the sample five-numbers summary: the smallest observation (sample minimum), lower quartile (Q_1), median (Q_2), upper quartile (Q_3) and the largest observation (sample maximum). Figure 4 shows a box- and-whisker plot of ozone pollution data across years. The vertical (box) variability indicates the variation of the annualized ozone pollution between the 20 geographic locations. A similar trend of increased pollution variability in the 2000's can be observed here as we saw in the spider chart on Figure 3.

Motion-Charts present a new paradigm for understanding multivariate data [15]. Figure 5 shows one instance of a motion chart of the ozone pollution data where year, the location coordinates (longitude and latitude), high-cover, annual and location-label are mapped onto play-time, X axis, Y axis, size, colour and category label, respectfully. The dynamic nature of Motion Charts enables the users to see the temporal changes of the annual ozone pollution rates for all locations, which gives clues to address the first driving motivational problem in this case study. These screencast videos (http://wiki.stat.ucla.edu/socr/index.php/SOCR_Videos_MotionCharts) may provide additional help with the utilization and functionality cues of the SOCR Motion Charts.

3.2. Quantitative data analyses

To gain a better understanding of the processes of interest, quantitative data analyses (QDAs) are typically employed after the initial visual and summarative EDA analyses. Here we demonstrate 3 rudimentary types of commonly used QDAs using SOCR Analyses, http://socr.ucla.edu/htmls/SOCR_Analyses.html [16]. These include significance testing, regression and ANOVA.

Hypothesis testing (or significance testing) is a method of making statistical decisions based on data. Statistically significant results of a hypothesis test indicate that the observed discrepancy between a model (null hypothesis) and the data is unlikely to have occurred solely by chance. Figure 6 shows the results of two hypothesis tests of whether the mean (for the parametric test) or the median (for the non-parametric test) annual ozone pollution rates of Los Angeles (location 2899) and San Francisco (location 2125) are different. In this case, the *null hypothesis* (model) would be that the two locations have the same ozone pollution rate (subject to random variation, or noise, in the data). Whereas the alternative research

hypothesis may be that the ozone rates at these two locations are actually distinct (i.e. geographic effect on pollution). If μ_{LA} and μ_{SF} represent the mean (or median, for the non-parametric test) annual ozone pollution rates for Los Angeles and San Francisco, respectively, then the null and research hypotheses are $H_0: \mu_{LA} = \mu_{SF}$, and $H_1: \mu_{LA} \neq \mu_{SF}$. Assuming the annual ozone pollution rates are independent with time, we can apply parametric (e.g. paired Student's t) or non-parametric (e.g. Wilcoxon signed-rank) statistical tests [16] to assess the strength of the evidence in the data to reject the null hypothesis (H_0) and therefore accept the alternative (H_1). The natural time-pairing of the 27 annualized ozone pollution measurements between the 2 locations indicates that a paired test is appropriate. Note that in this case, the parametric t -test ($p = 0.0088$) and the non-parametric Wilcoxon test ($p = 0.38709$) provide contradictory results. The parametric test is certainly more powerful; however, its application is dependent on parametric assumptions. Such parametric requirements may be validated using a variety of quantitative methods (e.g. SOCR Kolmogorov–Smirnov test) or qualitative (e.g. SOCR quantile–quantile (QQ) probability plot) techniques [14].

Regression analysis is a technique for modelling several variables (two or more) and analyzing the relationships between them. Typically, one of the variables is called a *dependent* (response) variable and the other one(s) are called *independent* (predictor, explanatory, regressor, covariate) variable(s). The inter-variable relations determined by regression analysis illustrate how the dependent variable changes when (some of) the independent variables vary. Regression analysis estimates a function (e.g. linear regression model function) of the independent variables to characterize the variation of the dependent variable. Regression analysis may be used for inference, prediction and forecasting. However, regression analysis may or may not imply (direct) causal relationships between the independent and dependent variables. In this case, the first driving motivational question relates to the temporal changes in the ozone pollution in California. A simple, first look at this relation may be obtained by fitting a linear regression model with Ozone-Pollution and Time, as the dependent and independent variables, respectively. Figure 7 shows the results of this simple linear regression analysis using SOCR Analyses [16]. Although this simple linear regression model cannot be expected to explain the entire variability in these ozone pollution data, it provides a clue to the trend of the pollution with time. Note that this is a *global* model including all geographic locations and the inference from this general model may not be applicable for individual locations. It may be more appropriate to fit regression model to each of the 20 individual locations separately. In this situation, it's possible that several of the linear models may show a steady decrease or a monotonic increase of ozone pollution with time. This location-specific effect may be washed out if we fit a cumulative global regression model for all locations.

ANOVA is another statistical technique where the observed data variance is partitioned into independent components based on the sources of the variation. ANOVA allows (parametric) statistical tests of whether the means of several groups are distinct. It can be viewed as a generalization of the Student's t -test to several groups and together with regression analysis is part of the general linear model [17]. ANOVA has an advantage over repeated uses of two-sample t -tests, as the latter approach increases the chance of type I (false positive) error. Suppose we are interested in testing a null hypothesis whether the (mean) annualized ozone pollution rates at all 20 locations are equal vs. an alternative that at least one pair of sites have distinct annual ozone pollution rates. Figure 8 shows the corresponding SOCR ANOVA analysis on the data clearly indicating that there are statistically significant geographic effects on measured ozone pollution. Of course, this may not be a causative effect, as there may be many other factors that play roles in this process (e.g. elevation of the measuring site).

Figure 9 shows the results of a two-way ANOVA where the *location* and *time* are used as factors and the ozone pollution measurement is the dependent (response) variable. Two-way ANOVA [14,18] allows the simultaneous study of the effects of two different factors on one univariate variable. In this case, both the main effects of each of the two factors (location and time), which represent the differences in means over levels of one factor collapsed over the levels of the other factor, are statistically non-trivial. From the standard ANOVA results table shown on Figure 9, both factors are significant with probability values below 10^{-5} . The R^2 and the QQ-probability plot of the residuals indicate that this linear model may not explain the complete variability in the data and may be improved by non-linear models and/or by including additional variables.

4. Conclusion

This article presents an example of a driving motivational problem that may be useful in multidisciplinary science courses. It demonstrates a complete hands-on investigative pipeline starting with a problem definition, data selection, modelling, interrogation, analysis and ending with statistical inference using California ozone pollution data.

Two specific ozone pollution questions were investigated in this case study. Only freely available data, resources and tools were utilized throughout this article to demonstrate formulation of useful research questions, identification and handling of appropriate datasets, use of appropriate technology for data manipulation and interrogation and statistical inference. The graphical results and quantitative findings from this case study are likely to generate more follow up questions than the original inquiries that motivated the study on the first place. Examples of questions that may arise from this initial inquiry into this problem include Does altitude have an effect on ozone pollution? Are there seasonal (monthly) effects on ozone pollution? Is there evidence of causality between different variables, or hidden variables which may play role in ozone pollution? The focus of this article was to demonstrate a real research driven case study with many variables, unknowns and interactions. This study requires explorative and quantitative analyses and demonstrates the use of some open-access resources for scientific data manipulation.

In addition to the EDAs and QDAs we demonstrated above, instructors and students may interrogate the ozone pollution dataset with other graphical and analytic methods. For instance, instructors may guide students to explore the possibility of autocorrelations in the time sequence over the 27 annual readings for each of the 20 different ozone-pollution recording sites. One may also employ SOCR Modeler [19] to fit probability distribution models to the ozone pollution data. Figure 10 illustrates one example of fitting a 3-term Gaussian mixture model to the annual ozone pollution values. SOCR Modeler (http://socr.ucla.edu/htmls/SOCR_Modeler.html) allows us to estimate the weights (mixture parameters) and the distribution parameters (means and variances) for all 3 Gaussian models that collectively describe the annual ozone pollution data. A Kolmogorov–Smirnov test comparing the distribution of the ozone pollution data and the automatically obtained Gaussian mixture model distribution indicates that the model fits the data well (p -value = 0.4579) – the results of this model is reported in the Results tab of the applet. Having a model estimate enables us to both simulate more ozone pollution data as well as do model-based (as opposed to data-driven) inference on this process.

Acknowledgments

This study was partially supported by the National Science Foundation, grants 0716055, 1023115 and 0442992, and by the National Institutes of Health, grant U54 RR021813. We are also thankful to the numerous users, students, collaborators and instructors who have provided feedback, critiques and ideas about improvements of the SOCR resources in the past several years. Valuable contributions from Dexter Whittinghill improved the readability of this

article. The critiques of the International Journal of Mathematical Education in Science and Technology editors also enhanced the manuscript.

References

- [1]. Franklin, C.; Kader, G.; Mewborn, D.; Moreno, J.; Peck, R.; Perry, M.; Scheaffer, R. Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report. American Statistical Association; Alexandria, VA: 2007.
- [2]. Holbrook J, Rannikmae M. The nature of science education for enhancing scientific literacy. *Int. J. Sci. Educ.* 2007; 29:1347–1362.
- [3]. Jordan J, Haines B. The role of statistics educators in the quantitative literacy movement. *J. Stat. Educ.* 2006; 14(2):1–12.
- [4]. Lunsford M, Fink A. Water taste test data. *J. Stat. Educ.* 2010; 18(1):1–19.
- [5]. Shelley, M.; Yore, L.; Hand, B.; Saul, E.; Hand, B. New Directions in Science Literacy Education, Quality Research in Literacy and Science Education. Springer; The Netherlands: 2009. p. 245-260.
- [6]. Neidell MJ. Air pollution, health, and socio-economic status: The effect of outdoor air quality on childhood asthma. *J. Health Econ.* 2004; 23:1209–1236. [PubMed: 15556243]
- [7]. McConnell R, Berhane K, Gilliland F, London S, Vora H, Avol E, Gauderman WJ, Margolis HG, Lurmann F, Thomas DC, Peters JM. Air pollution and bronchitic symptoms in Southern California children with asthma. *Environ. Health Perspect.* 1999; 107:757–760. [PubMed: 10464077]
- [8]. Islam T, McConnell R, Gauderman WJ, Avol E, Peters JM, Gilliland FD. Ozone, oxidant defense genes, and risk of asthma during adolescence. *Am. J. Respir. Crit. Care Med.* 2008; 177:388–395. [PubMed: 18048809]
- [9]. Islam T, Gauderman WJ, Berhane K, McConnell R, Avol E, Peters JM, Gilliland FD. Relationship between air pollution, lung function and asthma in adolescents. *Thorax.* 2007; 62:957–963. [PubMed: 17517830]
- [10]. McDonnell WF, Abbey DE, Nishino N, Lebowitz MD. Long-term ambient ozone concentration and the incidence of asthma in nonsmoking adults: The ahsnog study. *Environ. Res.* 1999; 80:110–121. [PubMed: 10092402]
- [11]. de Gruijl F, Longstreth J, Norval M, Cullen A, Slaper H, Kripke M, Takizawag Y, van der Leun J. Health effects from stratospheric ozone depletion and interactions with climate change, *Photochem. Photobiol. Sci.* 2003; 2:16–28.
- [12]. Dinov I, Christou N. Statistics online computational resource for education. *Teach. Stat.* 2009; 31:49–51. [PubMed: 21297884]
- [13]. Dinov I, Sanchez J, Christou N. Pedagogical utilization and assessment of the statistics online computational resource in introductory probability and statistics courses. *Comput. Educ.* 2008; 50:284–300. [PubMed: 19750185]
- [14]. Che A, Cui J, Dinov I. SOCR analyses: Implementation and demonstration of a new graphical statistics educational toolkit. *J. Stat. Software.* 2009; 30:1–19.
- [15]. Al-Aziz J, Christou N, Dinov I. SOCR motion charts: An efficient, open-source, interactive and dynamic applet for visualizing longitudinal multivariate data. *J. Stat. Educ.* 2010; 18(3):1–29.
- [16]. Che A, Cui J, Dinov I. SOCR analyses - An instructional java web-based statistical analysis toolkit. *J. Online Learn. Teach.* 2009; 5:1–19. [PubMed: 21546994]
- [17]. MacCullagh, P.; Nelder, J. Generalized Linear Models. 2nd ed.. CRC Press; Boca Raton, FL: 1991.
- [18]. Rutherford, A. Introducing ANOVA and ANCOVA: A GLM Approach. Sage; London, England: 2001.
- [19]. Dinov I. Statistics online computational resource. *J. Stat. Software.* 2006; 16:1–16.

Figure 1.
Geographic locations of the 20 ozone pollution measurement stations.

Figure 2.
Outline of the ozone pollution case study design.

Figure 3.

A spider chart of ozone pollution data across years. The top panel shows the spider chart and the bottom panel displays some of the rudimentary summary statistics that are reported by the SOCR Charts for all 20 geographic locations.

Figure 4. SOCR box-and-whisker plot of ozone pollution data across years. The chart whiskers indicate the variability of the annual ozone pollution between locations (across the years) and triangle symbols on top depict outliers. The insert figure represents the complete, untruncated, graph and shows the significant variability in the ozone pollution data for 2004–2006.

Figure 5.

A static image of the dynamic SOCR motion chart of ozone pollution data. The reader is encouraged to view the animated gif file (http://wiki.stat.ucla.edu/socr/index.php/SOCR_MotionCharts_CAOzoneData), explore the online activity and experiment with the SOCR Motion-Charts applet to see the full capability of this EDA tool to visually render multivariate data. The mapping panel (top-right) allows the user to specify appropriate variable-to-glyph appearance assignments which impact the rendition and the play behaviour of the motion chart.

Figure 6
 . Significance testing analysis comparing the annualized ozone pollution rates of Los Angeles against those of San Francisco. Top and bottom panels show the paired parametric (Student's t) and non-parametric (Wilcoxon signed-rank) test results, respectively.

- Regression Line: $OzonePollution = -191.6095 + 0.09667 \times Year$
  - Intercept:
    - Parameter Estimate: -191.6095127656
    - Standard Error: 28.4291851368
    - T-Statistics: -6.7398876135
    - P-Value: 0.0000000000
  - Slope:
    - Parameter Estimate: 0.0966959040
    - Standard Error: 0.0142644095
    - T-Statistics: 6.7788228014
    - P-Value: 0.0000000000
- Correlation(YEAR, OzonePollution) = 0.2805211066

Figure 7.

SOCR simple regression model of the linear relation between ozone pollution and time. The top panel illustrates the visual trend and the prediction interval for the ozone pollution across time. The bottom panel shows the quantitative regression model estimation and significance of the coefficients, respectfully.

Figure 8. ANOVA analysis of the ozone pollution data comparing the annual ozone pollution between all 20 sites (assuming time independence).

Figure 9.

Two-way ANOVA analysis of the ozone pollution using location and time as independent factors. Both the year and location variables show significant main effects ($p < 10^{-5}$).

Model fit results

Mixture model 0:

Weight=0.8925925925925926
Mean=1.778609959663692
Variance=0.4763119660879026

Mixture model 1:

Weight=0.046296296296296294
Mean=81.43999542236328
Variance=22.086456518866687

Statistical differences between the Means of model(0) and model(1):
 $Z_0 = 6.3135$; p-value = $1.3633E-10$

Mixture Model 2:

Weight=0.061111111111111111
Mean = 93.2121064157197
Variance = 12.045921809416626

Statistical differences between the Means of model(1) and model(2):
 $Z_0 = 0.78528$; p-value = 0.21614

Intersection Point(s):

12.176105913147536
87.42295124232308

Kolmogorov-Smirnoff test

Data vs. Mixture-M

Hypotheses:

H_0 : Data and model distributions are similar

H_1 : Data and model distributions are distinct

KS D-Statistics = 0.115544554
Z-Statistics = 0.81702
p-value = 0.4579174

Figure 10.

Fitting a three-term Gaussian mixture model to the ozone pollution data using SOCR Modeler, which enables model-based inference and random sampling of ozone pollution data. The left panel illustrates the three steps in the model fitting process (data-entry, parameter estimation and graphical and textual result reporting). The right panel contains the complete results of the mixture model fit (mixture and distribution parameters, statistical tests comparing each pair of mixture model distributions and testing the global quality of the entire model). To enhance the visualization of the data distribution and the corresponding mixture distribution model fit, the ozone pollution data here were scaled an order of magnitude ($\times 10$).

Table 1

The AQI provides both qualitative (colour) and quantitative (ppm) ozone pollution guidelines that pertain to human health.

Air Quality Index for Ozone		
Concentration Range (ppm)	Air Quality Description	Cautionary Statements for Ozone
0.00 – 0.060 ppm	Good	No health impacts are expected
0.061 – 0.075 ppm	Moderate	Unusually sensitive people should consider limited prolonged outdoor exertion
0.076 – 0.104 ppm	Unhealthy for Sensitive Groups	Active children and adults, and people with respiratory conditions (e.g., asthma) should <u>limit prolonged outdoor exertion</u>
0.105 – 0.115 ppm	Unhealthy	Active children and adults, and people with respiratory conditions (e.g., asthma) should <u>avoid prolonged outdoor exertion</u> . Everyone else, especially children and elderly, should limit prolonged outdoor exertion
0.116 – 0.374 ppm	Very Unhealthy	Active children and adults, and people with respiratory conditions (e.g., asthma) should <u>avoid all outdoor exertion</u> . Everyone else, especially children and elderly, should limit outdoor exertion

Table 2

A fragment of the CA ozone pollution dataset. The complete dataset is available online.

Variable	Location	Year	M1	M2	...	M12	Annual	Number of days with complete data	High coverage	Number of sites with complete data	Latitude	Longitude	Elevation
OZMAX1HR	2008	1981	0.07	0.11		0.1	0.11	350	95	1	34.462223	-120.025833	24
OZMAX1HR	2008	1982	0.07	0.08		0.09	0.15	361	99	1	34.462223	-120.025833	24
OZMAX1HR	2008	1983	0.06	0.1		0.07	0.14	359	98	1	34.462223	-120.025833	24
OZMAX1HR	2008	1984	0.07	0.07		0.07	0.14	356	97	1	34.462223	-120.025833	24
OZMAX1HR	2008	1985	0.09	0.09		0.09	0.13	360	99	1	34.462223	-120.025833	24
OZMAX1HR	2008	1986	0.1	0.11		0.07	0.11	352	96	1	34.462223	-120.025833	24
OZMAX1HR	2008	1987	0.08	0.08		0.05	0.17	351	96	1	34.462223	-120.025833	24
OZMAX1HR	2008	1988	0.08	0.09		0.06	0.11	359	98	1	34.462223	-120.025833	24
...	...	...	...	...	...	...	...	...	...	...	...	...	...