APPENDIX XI

MATES V

FINAL REPORT

Monitoring Data Treatment Methodologies

Appendix XI

Monitoring Data Treatment Methodologies

Measuring pollutants at low concentrations is more difficult than measuring pollutants at higher concentrations. Occasionally, the concentrations are so low that they are below the method detection limit (MDL). When this happens, we are only confident that the concentration could be as low as zero or as high as the MDL and is probably somewhere in between those two values. However, we cannot give a specific estimate of the concentration with any confidence when it is below the MDL. Every observation has a corresponding MDL. Laboratory technologies typically improve over time, and more recent observations tend to have lower MDLs than older observations. For example, the MDLs in the MATES V data are generally much lower than the MDLs in the MATES II data, see Appendix IV. Data with observations below the MDL are common in environmental data [1] and occur throughout the MATES data. Data below the detection limit are referred to as "nondetects" while data at or above the MDL are referred to as "detects". Statistical methods are available to perform calculations on data that include nondetects, in order to draw appropriate conclusions regarding spatial or temporal trends.

As laboratory technologies have improved over time, the statistical methods for handling data with nondetects have also improved and the improved methods are becoming more widely used in the environmental sciences. The MATES V analyses follow the guidance provided in Singh et al. (2006) [2] and Helsel (2012) [1]. Singh et al., 2006 [2] is an in-depth U.S. EPA-commissioned report on the topic of handling environmental data below detection limits, the authors of which consulted Dennis Helsel, the author of multiple textbooks describing methods to handle environmental data with nondetects, including Helsel (2012) [1]. General guidance from Helsel (2012) for handling data with nondetects recommends not deleting or ignoring the data below the detection limit and avoiding substitution[1] (e.g., 0.5*MDL) [1]. The analysis methods combine information about the proportions of nondetects with the numerical values of the data at or above the detection limit(s) [1].

The analyses for MATES II, conducted in 2000, used 0.5*MDL substitution to handle nondetects [3, pp. ES-7]. This approach was quite common and was endorsed by the U.S. EPA at the time [4]. Consistent with another EPA report [5], the analyses for MATES III (2008) and MATES IV (2015) reported specific values for data between the MDL and the Limit of Detection (LoD) and reported data below the LoD as zero [6, pp. Appendix VI-1, 7, pp. Appendix IV-1]. We updated our statistical methods for the MATES V measurement data analysis to make use of advancements in the science that are becoming more widely used for handling environmental data with nondetects. To be able to make direct comparisons of pollutant concentrations over time, MATES II through IV data are being re-analyzed alongside the MATES V data using these improved statistical methods.

[1] Substitution is only recommended for averaging points in cases where all data points have the same MDL [1, p. xix].

Helsel (2012) outlines three broad approaches to handling data with nondetects: 1) Maximum Likelihood Estimation (MLE), 2) nonparametric methods with a single MDL (applying the highest MDL to all observations if there are multiple MDLs), or 3) nonparametric survival analysis methods [1]. The MLE methods require that the data fit an assumed distribution and either have a small percent of the data be nondetects or have outside knowledge with which to determine the distribution [1]. MLE methods have been shown to perform poorly for skewed data with sample sizes smaller than 70 [1, p. 65]. The MATES data does not consistently meet the requirements of the MLE methods, so the two nonparametric approaches, 2 and 3, are used in analyzing the MATES data.

Summary statistics were generally calculated using the Kaplan-Meier method with Efron's bias correction (from nonparametric survival analysis methods) since it is the most generally applicable of the methods presented in Helsel (2012) [1, p. 85] (See Figure 1). A minimum sample size (number of detects plus the number of nondetects) of 10 is required, otherwise no statistics are calculated [2, p. 91]. Mean concentrations were, in most cases, calculated using the Kaplan-Meier Mean (KM mean) equations in Section 3.11 of Singh et al. (2006) [2] with Efron's bias correction [1, pp. 74-75, 8, pp. 100, 118]. The first exception was when more than 80% of observations were nondetects. In this case, a single estimate of the mean cannot be made for risk calculations, and therefore, we report the percent of data above the maximum MDL instead of calculating an estimate of the mean [1, p. 93]. For the purposes of giving upper and lower bound estimates for the risk calculations, zero substitution and MDL substitutions were used to calculate classical means of concentrations for use in the risk calculations, analogous to the method mentioned in Helsel (2012) [1, p. 94]. The classical mean is used in the rare occurrence when all concentrations were identical because the algorithm in Section 3.11 of Singh et al. (2006) [2] breaks down if there is no variation in the data. This can occur when all concentrations are above the MDL and have the same value or when less than 80% of the data are nondetects and all detects have values equal to the MDL, both of which are rare occurrences. When all data are above their respective MDLs, the KM mean yields the same numerical value as the classical mean.

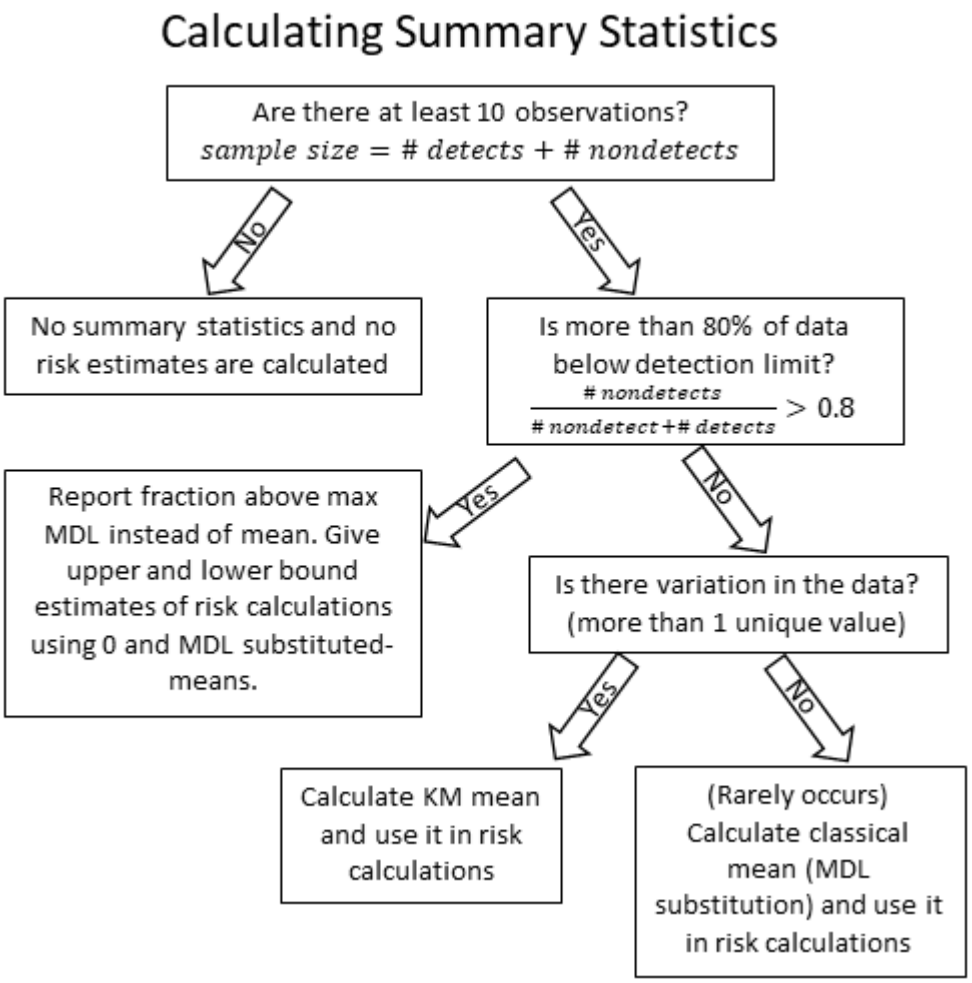

Figure XIV-1: Flow chart for determining how to calculate summary statistics and risk calculations for MATES data.

Calculations of confidence intervals follow guidance from Helsel (2012) [1] and Singh et al. (2006) [2]. Standard deviations and standard errors were calculated according to the equations in Singh et al. (2006) [2, pp. 31, 47]. The 95% confidence intervals were calculated using bootstrapping [1, pp. 103, 136-140]. Briefly, the KM mean is computed from a random sample of the data that is the same size as the data set. The random sampling is taken with replacement from the measurements, so that some measurements may be sampled multiple times while others may not have been sampled. This procedure is repeated 1000 times to give a distribution of KM mean estimates from 1000 random samples of the data. The 2.5^{th} and 97.5^{th} percentiles of the distribution of 1000 KM mean estimates provides the 95% confidence interval [1, pp. 103, 136-140]. The bootstrap 95% confidence intervals are only calculated if the data sample met the requirements to allow a KM mean to be calculated (See Figure 1). If a random sample had more than 80% of the data below the detection limit, then the KM mean cannot be calculated for that iteration and the classical mean using MDL substitution is used for that iteration instead of the KM mean. If none of the random samples used MDL substitution and the average of all of the KM mean estimates did not match the original non-boot-strapped KM mean within three

significant digits, then the bootstrap algorithm was run again with progressively larger number of random samples (up to a maximum of 50,000) until convergence was achieved, if possible. In the situation where the original data set had more than 80% below the detection limit and MDL and zero substitution were used to give upper and lower estimates as described in the paragraph above, bootstrapping was performed on the classical means for each the MDL and zero-substituted data sets to get the 95% confidence intervals for each.

For some MATES iterations (i.e., MATES II, III, IV, or V), some or all stations operated for more than a year. To calculate annual mean concentrations, the analysis was limited to data within the time periods shown in Table 1. MATES III was initially intended to collect observations during April 2004 through March 2005 and was extended for a second year due to heavy rainfall and concerns that the measurements would not represent typical meteorology. The MATES III final report presented annual averages for eight of the sites over the two-year monitoring period. Because the Huntington Park and Pico Rivera sites did not have a full second year of data, only data from the first year of measurements at these sites were used to calculate annual statistics [9, pp. ES-2, 10, pp. 1-1]. The current analysis uses the same averaging periods for each of the MATES III sites. In cases when there were multiple observations at a given station on a given day, the observations were merged by taking the (classical) mean of the replicate measurements prior to analyzing the data.

Table XIV-1: Date ranges for data included in this analysis.

MATES Iteration	**Start of data used**	**End of data used**
MATES II [11, pp. 1-2]	April 1998	March 1999
MATES III [9, pp. ES-2]	April 2004	March 2006
MATES IV [12, pp. Appendix X-1]	July 2012	June 2013
MATES V	May 2018	April 2019

References

[1] D. Helsel, Statistics for Censored Environmental Data Using Minitab and R, 2nd ed., Hoboken, New Jersey: John Wiley & Sons, Inc., 2012.

[2] A. Singh, R. Maichle, Lee and S. E, "On the Computation of a 95% Upper Confidence Limit of Unknown Population Mean Based Upon Data Sets with Below Detection Limit Observations," US EPA, Washington DC, 2006.

[3] South Coast Air Quality Management District, "Mulitple Air Toxics Exposure Study (MATES-II) Executive Summary," Diamond Bar, 2000.

[4] United States Environmental Protection Agency, "Data Quality Assessment: Statistical Methods for Practitioners EAP QA/G-9S," Washington, DC, 2006.

[5] EPA454/R-01-003, "Pilot City Air Toxics Measurements Summary," 2001.

[6] South Coast Air Quality Management District, "Appendix VI MATES III Final Report Summaries for the MATES III Fixed Monitoring Sites," Diamond Bar, 2008.

[7] South Coast Air Quality Management District, "Appendix IV MATES IV Final Report Summaries for the MATES IV Fixed Monitoring Sites," Diamond Bar, 2015.

[8] J. P. Klein and M. L. Moeschberger, Survival Analysis Techniques for Censored and Truncated Data, 2nd ed., New York: Springer, 2003.

[9] South Coast Air Quality Management District, "MATES III Executive Summary," Diamond Bar, 2008.

[10] South Coast Air Quality Management District, "MATES III Final Report Chapter 1 Introduction," Diamond Bar, 2008.

[11] South Coast Air Quality Management District, "Multiple Air Toxics Exposure Study (MATES-II) Final Report Chapter 1 Introduction," Diamond Bar, 2000.

[12] South Coast Air Quality Management District, "MATES IV Final Report Appendix X The Spatial and Temporal Trends of PM2.5, PM10, and TSP Components in the South Coast Air Basin," Diamond Bar, 2015.