4.0 OPTIONS FOR AUTOMATING COMPONENTS OF THE EMISSIONS FACTOR DEVELOPMENT AND DELIVERY PROCESS

4.1 Introduction

Traditionally, EPA has been the principal agency involved with the development and delivery of emissions factors in the US. However, the process of developing and delivering emissions factors involves numerous other organizations including source testing contractors; State, local, and Tribal agencies; private industry; and EPA contractors. While the various agencies and organizations involved with the development and delivery of emissions factors have automated aspects of the process, the automation has been limited and standardized automation procedures do not exist.

As a result of the limited automation technologies that have currently been applied to the emissions factor development and delivery process, the existing process is inefficient, time consuming, and expensive. For example, the processes of planning and conducting source tests as well as evaluating and delivering source test data often require redundant, manual entry of the same data into various software packages. In addition, the quality of source test data are often assessed by the testing contractor, the agency to whom the test data were submitted, and the organization tasked with calculating an emissions factor. This is a redundant and time consuming exercise. Furthermore, large numbers of source test reports and data are not easily accessible for use in the development of emissions factors.

EPA is interested in developing automation technologies to promote a more efficient and comprehensive emissions factors development and delivery process. Automating aspects of the process should:

- streamline the process,
- remove redundancies,
- allow additional data to be accessed in a cost effective and timely manner,
- provide standardization,
- capture appropriate process data so that emissions factors may be calculated,
- allow the subjective quality assessment of source test data and emissions factor ratings to be replaced with quantitative methods that characterize the data in statistical terms of precision and variation,
- improve the robustness of existing emissions factors,
- allow for the development and continuous refinement of emissions factors with current data.

For example, the use of automation technologies could allow the EPA and others to collect data from all source tests conducted for State, local, and Tribal agencies in the United States. The universal availability of these basic data would allow their use for emissions factor development as well as for the development of Federal, State, and local rules; PSD/NSR control evaluations; and other programs where actual emissions information is needed.

Four components of the emissions factor development and delivery process could be improved through the use of automation. These components are:

1. The preparation of source test plans and the evaluation and delivery of source test data.
2. The storage and availability of source test data.
3. The analysis of source test data for the development of emissions factors.
4. The dissemination of emissions factors and related information.

MACTEC has developed options for developing automation technologies to improve each component of the emissions factor development and delivery process. The following sections document the existing automation technologies associated with each of the components identified above and options for EPA to consider implementing as a means to improve the process. Because the emissions factor development process hinges upon the availability of source test data, many of the options discussed herein focus on the automation of the source test planning and data evaluation, delivery, and storage steps.

4.2 Existing Emissions Factor Development and Delivery Automation Technologies

A review of existing automation technologies indicates that some automation technologies are currently used by EPA, State and local agencies, and source test contractors to complete the emissions factor development and delivery process. The automation technologies currently used to complete each of the components of the process are identified and described in the following sections.

4.2.1 Component 1: Source Test Planning, Evaluation, and Delivery

Source testing contractors typically submit source test protocols and plans to State, local, or Tribal agencies for review prior to initiating testing. The protocols/plans are typically developed in word processing programs and transmitted to the reviewing agencies in either hard copy or in Adobe's portable document format (PDF). Although eXtensible markup language (XML) may be used in Adobe's PDF software, XML is rarely used. As a result, testing contractors must manually incorporate the reviewing agencies' comments into their protocols/plans. Furthermore, if the reviewing agency desires to verify any calculations included in the test protocol/plan, the data must be manually entered into spreadsheet programs or calculated by hand.

Upon completion of a source test, most source test contractors use spreadsheets to automate the manipulation of the data gathered. Depending upon the source testing firm and pollutants being measured, the spreadsheets are populated by manually entering the data or by transferring the data electronically from a data acquisition system. The data from spreadsheets are typically transferred electronically to word processing software for the development of reports that are subsequently sent to reviewing agencies in either hard copy or PDF format. Again, XML is rarely used during the transfer of the source test reports; therefore, further electronic manipulation of the data contained in the reports is difficult and time consuming.

Federal, State, and local agencies often use spreadsheets to evaluate the quality and accuracy of source test reports. However, because the source test reports are typically transferred to the reviewing agencies in hard copy or PDF format, the test data must be manually entered into the spreadsheets prior to evaluation.

4.2.2 Component 2: Storage and Availability of Source Test Data

Many State, local, and Tribal agencies maintain source test protocols, plans, and reports as well as any quality assurance documentation in hard copy or PDF format only. Although the agencies often use spreadsheet or word processing programs to maintain indices of the source tests conducted, organizations that desire to develop emissions factors based upon the source test reports often must undergo a labor intensive, time consuming, and expensive process to identify the appropriate source test reports and obtain them.

4.2.3 Component 3: Analysis of Source Test Data for Emissions Factor Development

After source test reports have been obtained, organizations often use spreadsheets to evaluate the quality and accuracy of the reports. However, because the source test reports are typically obtained in hard copy or PDF format, the test data must be manually entered into the spreadsheets prior to evaluation.

Once the required data have been extracted from the test reports, emissions factors are typically calculated using automated methods such as spreadsheets and databases. The spreadsheets and databases also provide the capability to assess different scenarios for aggregating and segregating the data, and for identifying anomalies in the data.

4.2.4 Component 4: Dissemination of Emissions Factors and Related Information

Currently, EPA makes a fairly comprehensive list of air emissions factors available on its web site at http://www.epa.gov/ttn/chief/. At this site, users can access EPA's *Compilation of Emissions Factors*, commonly referred to as AP-42. The AP-42 sections are provided in PDF format and each section can be downloaded and printed. However, the emissions factors may not be efficiently transferred electronically from the PDF files to spreadsheets for use in air emissions inventory calculations or other analyses. Rather, the factors usually must be manually entered into the spreadsheets or databases for further use.

This same web site also provides access to two emissions databases, the Factor Information and Retrieval (FIRE) system and SPECIATE. Both of these tools allow users to search for emissions factors using various selection criteria and to download search results in database, comma separated variable, and/or text format. In addition to the aforementioned databases, the Chief web site provides access to several automated tools that are used to calculate emissions factors and/or emissions from specific source categories. These tools include the Landfill Gas Emissions Model, TANKS, and WATER9 programs. Similar tools for on-road

mobile sources (MOBILE6) and nonroad mobile sources (NONROAD) are available from the EPA's Office of Transportation and Air Quality at http://www.epa.gov/otaq/.

Various State agencies also provide mechanisms for retrieving emissions factors from their web sites. For example, the Michigan Department of Environmental Quality has an automated system for retrieving emissions factors called the Michigan Air Emissions Reporting System (MAERS), and the California Air Resources Board (CARB) has an automated system for retrieving toxics emissions factors called the California Air Toxics Emission Factor (CATEF) database. This system allows users to query a database that contains AP-42 emissions factors using the Internet. Nevertheless, no automation technologies were found in to be in use that fully integrate all facets of the emissions factors delivery process.

4.3 Options for the Development of Automation Technologies

This section describes automation technologies that might be developed and applied to improve each of the four components of the emissions factor development and delivery process. Three levels of automation ("low," "intermediate," and "high") are presented for each component; the options are summarized in Table 1. Advantages, disadvantages, and preliminary cost data for each option are presented in the following subsections.

4.3.1 Component 1: Source Test Planning, Evaluation, and Delivery

The tasks of source test planning, evaluation, and delivery could be subdivided further, but it is reasonable to group these tasks together as they are highly related. The lowest level of automation would involve no transmission of data from the source test contractors to State and local agencies. Therefore, State and local agencies would continue to receive source test reports and related information in hard copy format, or perhaps in Adobe PDF. The intermediate scenario would provide for data incorporation into an electronic document from which a user could copy data to be used in other software applications. Using the highest level of automation would streamline the very labor intensive processes associated with the paperwork transmission and transcription portions. Conceptually, data would be received in a data rich environment over the Internet or through automated web pages. Given high levels of automation, data from facilities and source testing firms would be received in a data rich environment (XML, HTML). Ideally, source tests would have standardized formats and data elements. The completed product would be on several documents which would be used to deliver data and assess the quality of the data.

There are advantages to automating this component of the emissions factors development process. The process of streamlining this component could not only save time, but also make the data much more useful for Components 3 and 4. Enabling a data rich environment from the initiation of the emissions factor development process would enable a more smooth flow of data to all aspects of this process. In addition, it would reduce the amount of time spent by personnel performing manual data entry.

TABLE 1. OPTIONS FOR AUTOMATING COMPONENTS OF THE EMISSIONS FACTOR DEVELOPMENT AND DELIVERY PROCESS

Level of Automation	Description
Component 1: Source Test Planning, Evaluation, and Delivery	
Low	No changes from present procedures would be implemented. Source testers would typically submit data to review agencies in hard copy or PDF format. After extracting data manually from source test documents, reviewing agencies would evaluate the data using manual methods or automated tools.
Intermediate	Standardized formats and data elements would be defined for source test reports. Source testers would typically submit data to review agencies in word processor and/or spreadsheet (i.e., electronic) format, but the data might not be in a data rich environment (e.g., XML, HTML, etc.). After extracting data electronically from source test documents, reviewing agencies would evaluate the data using automated tools.
High	Standardized formats and data elements would be defined for source test protocols, test reports, and quality assurance forms. Source testers would typically submit data to reviewing agencies in a data rich environment (e.g., XML, HTML, etc.). Reviewing agencies would evaluate source test documents using automation tools.
Component 2: Storage and Availability of Source Test Data	
Low	Source testing documents would be collected from State, local, and Tribal agencies by EPA or EPA-sponsored contractor personnel on a regular basis. The documents would be scanned and stored on CD ROMS or on a web server for public access.
Intermediate	EPA would receive source test data from State and local agencies in electronic format. EPA would request State and local agencies to provide data that meets specific criteria (i.e., graphic images data deficient but text based documents, preferably data rich text documents.)
High	Source testing documents (test protocols, test reports, and quality assurance forms) would be stored on State, local, and Tribal agencies' public servers. Depending on the format and storage method, differing levels of effort would be required to obtain the data. Data mining software could be used to search the web for publicly available data.
Component 3: Analysis of Source Test Data for Emissions Factor Development	
Low	EPA would use optical character recognition technology to capture applicable portions of the source test report and State/local agency source test assessment. Factor development would still entail manual labor to pool/aggregate data and generate factors for each broad category of process/control/pollutant. Alternatively, if set rules could be developed for pooling or aggregating data, this task could be automated.

TABLE 1. CONCLUDED

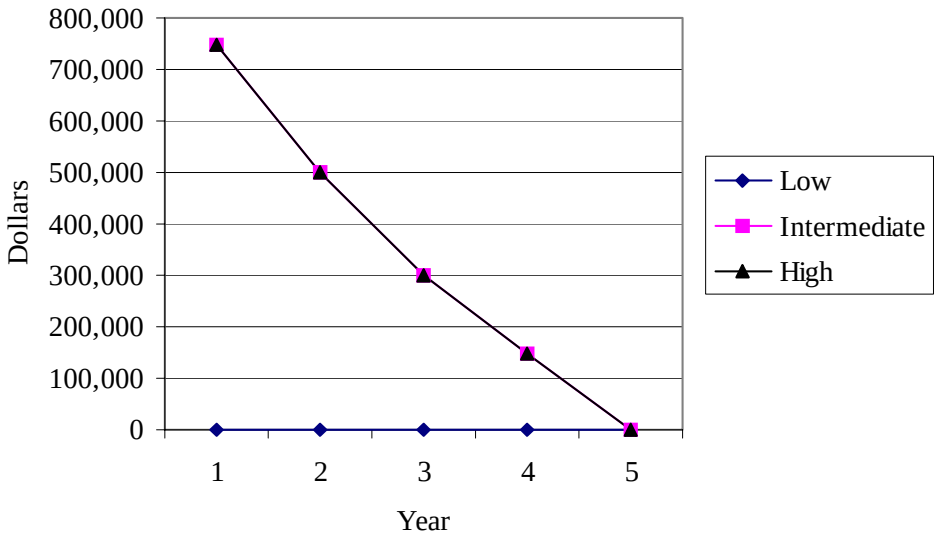
Level of Automation	Description
Component 3: Analysis of Source Test Data for Emissions Factor Development (continued)	
Intermediate	EPA would receive data in formats precluding the use of optical character recognition. Factor development would still entail manual labor to pool/aggregate data and generate factors for each broad category of process/control/pollutant. Alternatively, if set rules could be developed for pooling or aggregating data, this task could be automated.
High	Data rich text would be captured and used for the emissions factors development process. Factor development would still entail manual labor to pool/aggregate data and generate factors for each broad category of process/control/pollutant. Alternatively, if set rules could be developed for pooling or aggregating data, this task could be automated.
Component 4: Dissemination of Emissions Factors and Related Information	
Low	No changes from present procedures would be implemented. The current methodologies used to disseminate emissions factors (e.g., AP-42, FIRE, SPECIATE) would be retained.
Intermediate	Emissions factors and related information would be made available through an automated system similar to those used by stock trading and information web sites (e.g., www.morningstar.com). However, the ability to customize emissions factors to meet an end user's specific needs would not be provided.
High	Emissions factors would be made available using an automated system similar to those used by stock trading and information web sites (e.g., www.morningstar.com). The system would be developed so that the end user could tailor emissions factors to meet their specific needs: i.e., reflect criteria such as the age of the equipment tested, the size of equipment tested, the quality assessment values associated with the source test reports, the dates the tests were conducted, and the regional locations of the sources.

It is hard to conceptualize real disadvantages to the automation of this component of the emissions factors development process. The technology exists that allows for this process to be performed better and cheaper.

Figure 1 provides a graphical representation of the estimated 5-year costs for the low, intermediate, and high levels of automation. The estimates indicate that there are no costs associated with the low level of automation, and the highest costs are associated with the intermediate levels of automation. The data table from which the data for this figure were drawn are presented in Attachments 1-3 to this memorandum. The data tables are arranged by level of automation and include all tasks described in Table 1. Therefore, Attachment 1 present the costs associated with all tasks for the low automation option. Attachment 2 present the assumptions and costs associated with the intermediate level of automation for all tasks. Attachment 3

contains the assumptions and cost estimate for the high level of automation option. Attachment 4 contains the data and calculations used to determine the approximate number of source tests conducted annually in the US.

Figure 1. Costs to EPA for Source Test Planning, Evaluation, and Delivery



4.3.2 Component 2: Storage and Availability of Source Test Data

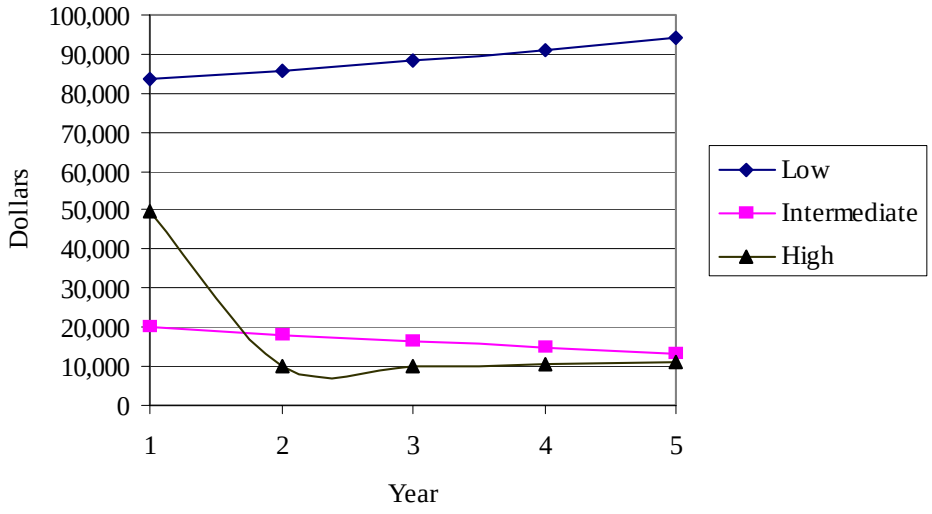
Options within this component range from visiting State and local agencies to scan test reports, permit files, and other associated information to the enablement of data mining of State and local agency public servers by EPA. At the lowest level of automation, EPA would visit State and local agencies to scan their source test reports and store these on CD ROMs or other data storage devices. In addition to collecting source test reports, EPA could potentially collect information pertaining to all aspects of the quality assured report, which might also include State and local agency evaluations. At the intermediate level of automation, EPA would receive source test reports from State and local agencies in an electronic format. EPA would either request State and local agencies to submit test data or provide a means for State and local agencies to provide data that met specific criteria. For example, EPA could specify the file format, or the file contents. Finally, high levels of automation would entail State and local agencies posting their source test reports and related information on public servers that would be "mined" using EPA software. The data mining software would not need to be unique to EPA, but could potentially be used by anyone that wanted information that is more up to date than EPA's data.

The lowest level of automation would be both very time consuming and labor intensive. In addition, it is questionable whether EPA would be able to collect all of the necessary data from State and local agencies. Cooperation from State and local agencies would be required to access all the files and scan them. Finally, the EPA would not be putting in place a mechanism to obtain the most recent source test reports. Therefore, it is conceivable that EPA would have to repeat this data collection exercise in given time intervals (e.g., five year, ten year intervals). This would preclude the agency from developing the most up to date emissions factors.

For the reasons listed above, it would be beneficial for the EPA to develop a system that allows it to obtain more current source test reports and associated information. The intermediate and higher levels of automation would enable this process. Both the intermediate and high levels of automation would ensure that EPA would receive current source test data. Unfortunately, using intermediate to high levels of automation would probably preclude the EPA from receiving historical source test reports as it is doubtful that State and local agencies would have the resources to transfer them to an electronic format. The lack of data from previous years' source tests would preclude their use in emissions factors development efforts.

Figure 2 contains a graphic representation of the 5-year costs associated with this task. As with the previous task, the data tables used to estimate these costs are presented in Attachments 1-3 to this memorandum. The data table also includes all the assumptions that were made to generate the cost estimates.

Figure 2. EPA Costs Associated with the Storage and Availability of Source Test Data



4.3.3 Component 3: Analysis of Source Test Data for Emissions Factor Development

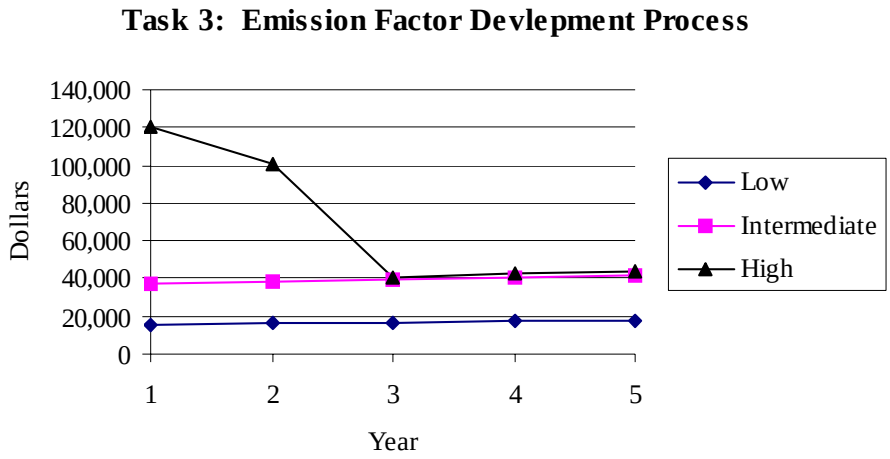
Much of the level of effort required for this component depends on the upstream components that are selected and implemented. For example, using low levels of automation entails the use of optical character recognition software to scan in applicable portions of the source test report and related information to translate the process, control, emissions, and quality assessment of the source test. Using intermediate and high levels of automation would not require that EPA scan the documents as they would be received in electronic format. The ideal goal would be to capture the appropriate information from data rich files (source tests reports and assessments). This would enable the EPA to spend the least amount of resources in the development of emissions factors.

In each of these situations, the factor development may be by a manual assessment to pool data and generate factors for each category of process/control/pollutant. Therefore, this process would still require the use of custom spreadsheets, or possibly databases. Alternatively, if set rules could be established for pooling or disaggregating the data, emissions factors development could be performed by software and allow one person to review and publish the resulting emissions factors assessment.

Receiving electronic files from states and local agencies will significantly cut down the amount of time required to process these data. The alternative, which is to use optical character recognition to make data within files manipulable, is both expensive, cumbersome, and time consuming. Therefore, it is recommended that EPA pursue a course by which it will receive files in an electronic format that lends itself to further data processing.

Figure 3 contains the 5-year cost estimate for low, intermediate, and high costs associated with the emissions factor development process. The data tables and assumptions used to estimate these costs are presented in Attachments 1-3 to this memorandum. Furthermore, many of the cost elements for this task were obtained from the Task 3 memorandum which was submitted earlier to EPA.

Figure 3. EPA Costs Associated with the Analysis of Source Test Data for Emissions Factors Development



4.3.4 Component 4: Dissemination of Emissions Factors and Related Information

EPA currently makes databases and programs available over the Internet that deliver and calculate emissions factors. This system, albeit functional, can be very cumbersome. For example, individual databases and programs have to be downloaded in order to obtain emissions factors. In some instances, it has been shown that these programs can be configured or altered to run over the Internet. For example, Michigan's MAERS, and CARB's CATEF databases can be queried online. It is highly unlikely that programs that require input files (such as MOBILE6) can be configured to run over a web server.

The lowest level of automation would entail maintaining the current system of models, processors, and PDF documents made available on EPA's web site. The current EPA web site could be streamlined to allow users to access the various programs through a single web site similar to those used by financial companies (i.e., Morningstar, Hoovers, Standards & Poors, etc.). Intermediate levels of automation would entail more enhanced web sites that offer more features and functions. For example, rather than having to download FIRE, a user could query an online database similar to MAERS. High levels of automation would allow a user to generate emissions factors online based upon such criteria as the quality assessment values of the source test reports, the date the test was conducted, the regional location of the source, and other parameters that may have an effect on the emissions factors.

Private businesses are successfully using the world wide web to promote their businesses, and to disseminate information. The web sites mentioned above are an indication of the power of these automation technologies. These same technologies can be applied to the emissions

factors distribution process. By making the information more user friendly and accessible, EPA would facilitate their use, and potentially promote better use of existing information. A potential disadvantage to this reconfiguration would be cost.

Figure 4 presents the 5-year costs associated with the distribution of emissions factor information for the low, intermediate, and high levels of automation. The cost estimates are based on MACTEC's experience in providing these type of services to EPA. In addition, the cost estimate for the maintenance of the FIRE database is based on MACTEC's seven year track record in maintaining this system for EFIG.

4.3.5 Discussion of Costs

Figure 5 provides a comparison of the total five-year costs for all components by level of automation. As is evident from Figure 5, the initial costs for implementing the intermediate and high levels of automation are significantly higher than the low level of automation. The costs for the intermediate and high levels of automation are higher due to the need to disburse grant monies to States, and due to the development of a powerful and interactive web site. The costs for the higher levels of automation are projected to fall below the costs for the low level of automation after year 4. Following year 4, it is anticipated that no more grant monies will be disbursed to States, and that the development of an enhanced web site will be complete.

Figure 6 provides a bar chart that shows all costs stacked in individual bars by level of automation and year. The bars are arranged by level of automation and year. Therefore, the year 1 costs of the low level of automation are presented first. Each bar is stacked with the costs for the four components that were identified as being fundamental to the automation of the emissions factors development and delivery process. For the low level of automation, the principal costs are split roughly evenly between the distribution of emission factor information and source test planning, evaluation, and delivery components. The grant monies disbursed during the first four years of the program account for the highest costs associated with the intermediate and high costs of automation. After the initial four years, the principal costs for these options are the distribution of emission factor information and the emission factor development process.

Figure 7 provides an estimate of the number of emission factors that may be developed using the varying levels of automation. The low level of automation is projected to yield the least amount of data that can be used for the development of emissions factors, and consequently the least number of emissions factors. The intermediate level of automation is projected to provide sufficient data for the development of over 500 emission factor per year. No increase in the amount of data collected for the development of emissions factors is projected due to the mix in levels of automation and formats of incoming source test reports. Finally, it is projected that the use of high levels of automation will allow for an annual increase in the collection of data used for the development of emissions factors. Consequently, this will enable the development of more emissions factors, which is reflected in Figure 7.

Figure 4. EPA Costs Associated with the Dissemination of Emissions Factor and Related Information

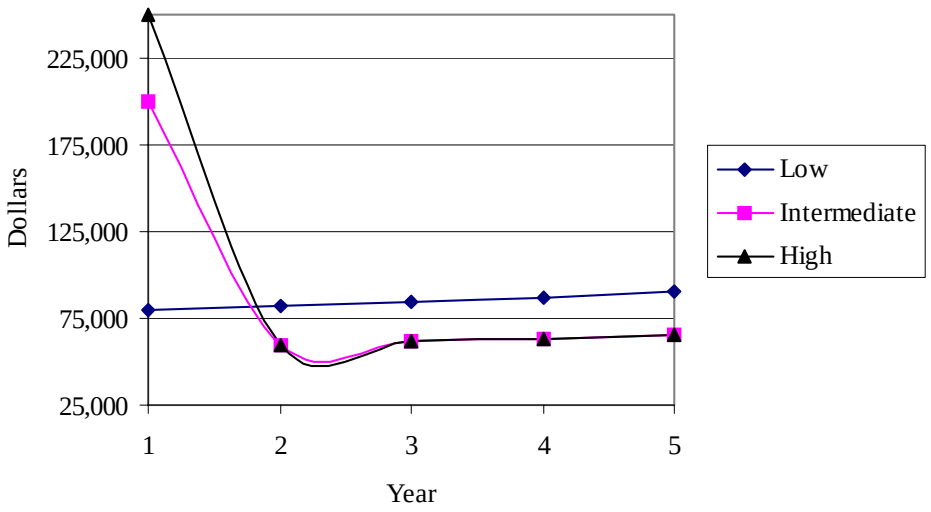


Figure 5. Total Cost to EPA for All Components

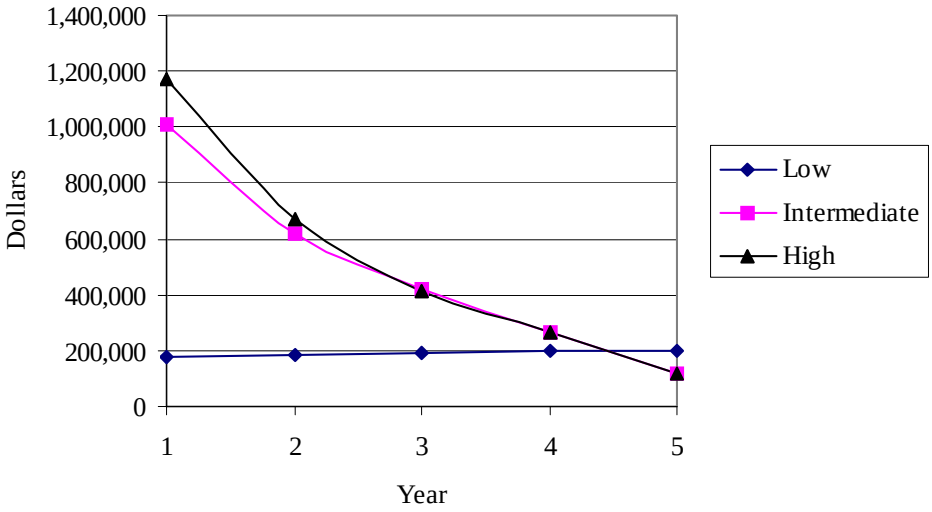


Figure 6. Comparison of EPA Costs for Automation Options

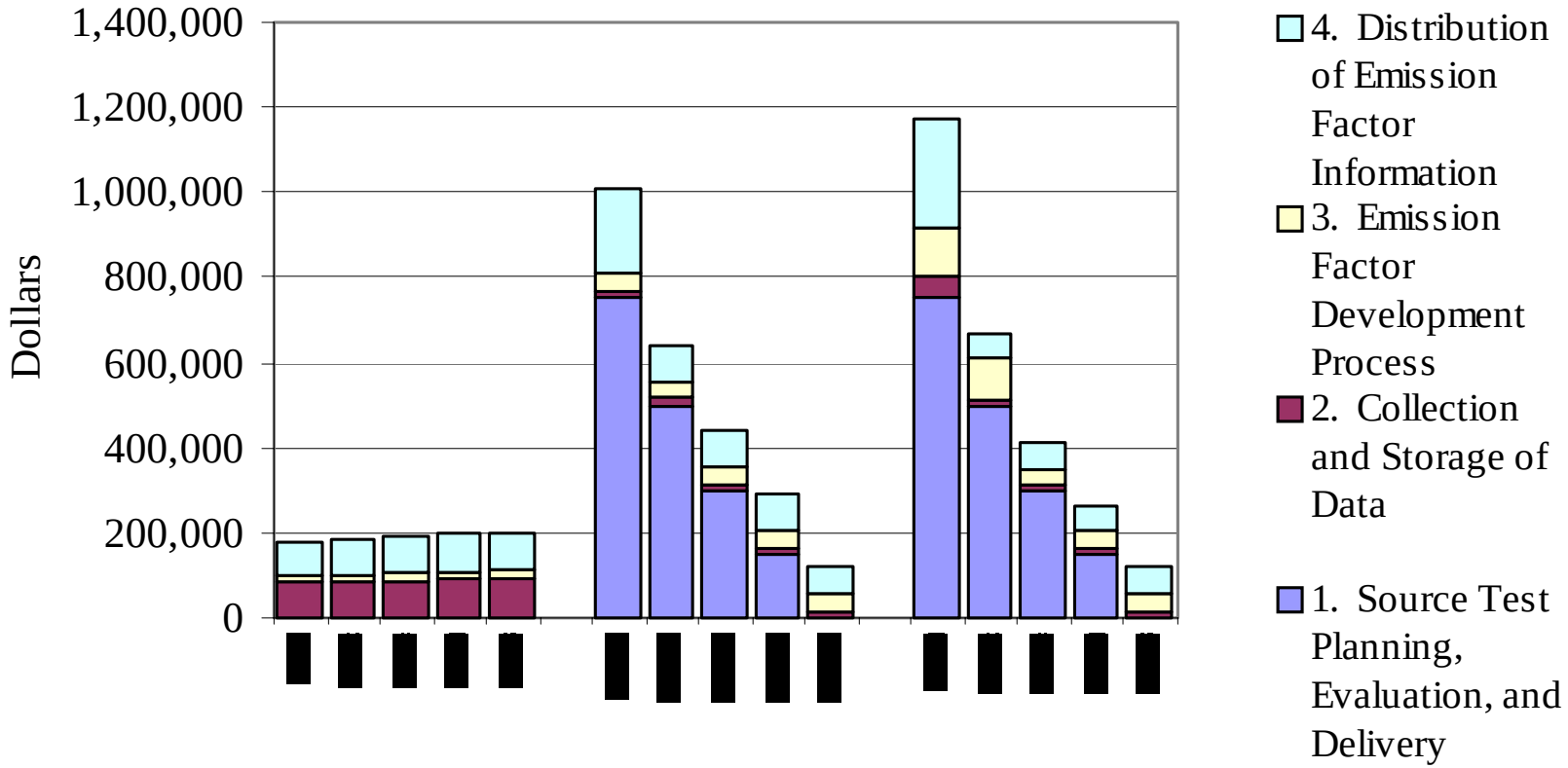
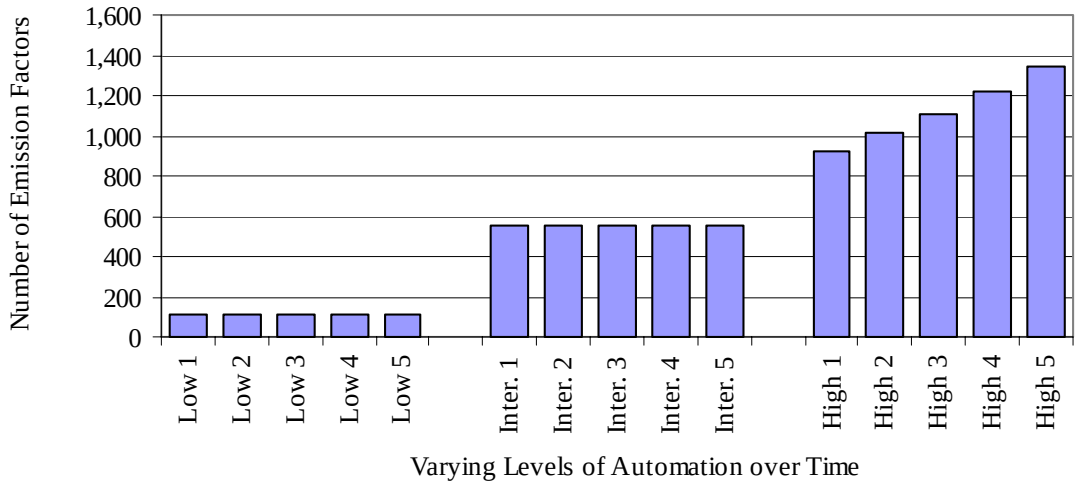


Figure 7. Emission Factors Available for AP-42 Development



4.4 Conclusion

EPA's current emission factor program has been found to be inefficient, time consuming, and expensive. The use of automation technologies would enable EPA to collect data and develop emissions factors more efficiently. As is shown in Figure 5, the initial costs for the high and intermediate automation options are higher than for the low automation option. After year 4, the costs for the higher levels of automation drop below the costs for the low level of automation option. In addition, as is shown in Figure 7, the number of emissions factors that can be developed from the data collected using the low level of automation remains stagnant whereas the use of higher levels of automation increases the amount of data collected and therefore increases the number of emissions factors that can be developed.