Submitted to:

PENNSYLVANIA DEPARTMENT OF TRANSPORTATION

USING SPATIAL TOOLS TO ANALYZE CRASH AND ROADWAY DATA PROJECT

RFQ No. 06-05 (C08)

FINAL REPORT

Draft Final Report on a research project to define a methodology that will help PennDOT "take highway safety to the next level".

Prepared by:

GeoDecisions®
A Division of Gannett Fleming

February 25, 2008

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
FHWA-PA-2008-002-060508		

4. Title and Subtitle	5. Report Date
Using Spatial Tools to Analyze Crash and Roadway Data, Project Number 0605(08)	February 25, 2008
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Gannett Fleming, Inc. (J. Cichocki, A. Sarvis)	GF Project Number 048026

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
Gannett Fleming, Inc. PO Box 67100 Harrisburg, PA 17106-7100	11. Contract or Grant No. 355I01-060508

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
The Pennsylvania Department of Transportation Bureau of Planning and Research Commonwealth Keystone Building 400 North Street, 6th Floor Harrisburg, PA 17120-0064	Final Report February 2007 – February 2008
	14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

PennDOT engaged Gannett Fleming to conduct research into best practices in the use of geospatial analysis tools for highway safety analyses. The goals of the effort were to define a methodology for PennDOT to follow in identifying the best candidate locations for highway safety improvements, and to develop a Proof of Concept to test the proposed methodology.

After conducting interviews and workshops involving more than 35 of PennDOT's stakeholders in highway safety processes, Gannett Fleming interviewed highway safety managers in five other state and federal highway agencies to determine what innovative tools and practices are currently being used. Gannett Fleming 's research also included a review of literature related to the study from more than 80 sources.

Based on Gannett Fleming's research and analysis, PennDOT selected the "Highway Safety Data Relationships Knowledge Base" for further research. The knowledge base is an information repository based on concepts in data mining and expert systems. It uses advanced statistical analysis methods and expert business knowledge rules to discover data patterns based on correlation and other forms of relationships in the data. The knowledge base can be applied to diagnosing specific combinations of data attributes and features that may indicate the causative factors among homogeneous populations of crashes.

Most highway safety data analyses involve studying correlations among multiple data sets. The knowledge base is an innovative and comprehensive tool for such an application. It provides a framework for identifying and managing relationships among many combinations of data sets that are useful in highway safety analyses.

Gannett Fleming proceeded to develop a prototype as a proof of concept. Gannett Fleming demonstrated the prototype using actual PennDOT crash data. Three analysis scenarios were demonstrated: evaluating safety programming alternatives for alcohol-involved crashes, diagnosing data patterns of crashes at a selected highway location, identifying potential sites for system-wide deployment of a selected countermeasure.

17. Key Words		18. Distribution Statement
Highway safety	Data mining	No restrictions. This document is available from the National Technical Information Service, Springfield, VA 22161
Crash data analysis	Knowledge base	
Safety Management	Data pattern recognition	
Spatial analysis	Statistical correlation	
Geospatial analysis	K-means cluster analysis	

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	25	

Table of Contents

1 Introduction

1.1 Document Purpose

This document is the deliverable report for Task E: Final Project Report in the research project titled "Using Spatial Tools to Analyze Crash and Roadway Data" (RFQ Number 06-05 (C08). This document provides a summary of the data inventory, interviews/research, and methodology selection phases of the project and provides details and results from the Task D proof of concept. Based upon the results of this project, PennDOT's vision for the future of the proof of concept is also presented.

In many ways, the Pennsylvania Department of Transportation (PennDOT) has been a national leader in the use of information technologies for highway safety data analysis. The Bureau of Highway Safety and Traffic Engineering (BHSTE) has continued to aggressively pursue new approaches to applying technology tools in its crash reduction goals. In 2006 PennDOT authorized a project to extend its capabilities in the application of geospatial information technology for crash data analysis. GeoDecisions was awarded the contract to execute the "Using Spatial Tools to Analyze Crash and Roadway Data" project as part of a larger program conceived by the Safety Management Division in BHSTE to "take safety to the next level".

The purpose of this research project was to explore new methods of applying geospatial technologies to analyze crash and roadway data producing meaningful information to support Pennsylvania's highway safety goals. The scope of work included performing research through literature review and interviews of other transportation agencies to determine what innovative tools and practices are currently being utilized for the same reasons that PennDOT was seeking. The project scope also included defining a methodology for PennDOT to follow in identifying the best highway locations for safety improvements, and testing that methodology via proof of concept.

After completing the research, GeoDecisions identified several alternative methodologies, from which PennDOT selected the Highway Safety Data Relationships Knowledge Base. GeoDecisions tested the Knowledge Base by developing a prototype for proof of concept. GeoDecisions demonstrated the proof of concept using PennDOT's crash data. The demonstration pointed out how the Knowledge Base met the criteria established earlier during the research phase of the project. This document is a report on the project performance.

1.2 Document Organization

This document contains four sections, plus appendices:

- 1. Introduction – Defines the purpose of the document, and provides an overview of the document contents and version information.

- 2. Final Report Context and Background– Presents the primary drivers and background for the project, previous project tasks (Tasks A – C), and an overview of the Task D proof of concept.
- 3. Proof of Concept Defined – Presents a review of the decision rationale for choosing the Knowledge Base Proof of Concept and an overview of what a knowledge base is.
- 4. Proof of Concept Results and Recommendations – Presents an analysis of the Proof of Concept results and recommendations for extended implementation at PennDOT.

1.3 Version Information

Version Num.	Edit Date	Edited By	Comments
0	November 30, 2007	J. Cichocki, A. Sarvis	Preliminary Outline
1.0	February 8, 2008	J. Cichocki, A. Sarvis	Draft Final
1.0	February 25, 2008	J. Cichocki	Approved Final document

2 Final Report Context and Background

2.1 Statement of Problem

The Pennsylvania Department of Transportation (PennDOT) is responsible for continually maintaining and improving the safety of the Commonwealth's transportation network. Addressing this responsibility requires successful development and implementation of processes to: identify and analyze the locations and contributing factors of crashes; select locations that have the highest potential for improvement; evaluate possible countermeasures and their probable impact on safety; and track the effectiveness of implemented countermeasures. The tangible and measurable success of this goal are stated clearly in PennDOT's Comprehensive Strategic Highway Safety Improvement Plan (CSHSIP) which calls for reducing traffic fatalities in the Commonwealth to 1.0 per 100 million vehicle miles traveled by 2008.

The perpetual need for improved processes and tools means that PennDOT must constantly look at industry trends and best practices for new opportunities. Modern geographic information systems (GIS) and database software show great potential for integrating and analyzing crash and roadway data. PennDOT has used its existing crash location clustering algorithms for many years now, but is specifically interested in identifying new spatial analysis tools that can be used to support its safety improvement processes.

PennDOT does have an extensive set of established and effective crash analysis tools and methodologies including, crash databases, the spatial query tool CDART, algorithms for developing Location Priority Lists, and customized file of standard engineering countermeasures. In order to best address the problem of highway safety, PennDOT needs to leverage these existing tools and datasets, including spatial data/tools, to provide better information for decision making. All of these existing datasets, tools and processes are core components for crash location analysis at PennDOT, but each provides a foundation and opportunity for potential improvement through the use of new "state of the practice" tools.

2.2 Project Objectives

In light of the need for continual highway safety improvements, and in recognition of the new compliance requirements of SAFETEA-LU, PennDOT entered into this project to investigate new methods to improve crash data analysis capabilities and produce meaningful information to support the States highway safety goals.

The project was organized into multiple tasks intended to methodically gather requirements and expectations, and research "state of the practice" spatial tools for highway safety analysis. Comprehensive research began with a review of PennDOT's safety analysis goals; current crash and roadway data; and current crash analysis processes. It also included research into the processes being used by other state and

federal government agencies and to assess the capabilities available through commercial-off-the-shelf (COTS) GIS and spatial database products. The information gathered during these tasks was intended to establish candidate tools for an improved approach for identifying locations where safety improvements are needed. From these candidates a tool was selected and then tested by conducting a Proof of Concept analyses. This organized approach for discovering new highway crash analysis tools is detailed in each of the sections below.

2.3 Task A Analysis

A review of PennDOT's current safety Analysis capabilities and future expectations established the technology/data foundation and comprehensive safety analysis themes that were of significant interest to PennDOT's Bureau of Highway Safety and Traffic Engineering staff.

The first project task was intended to capture the current status of PennDOT's safety analysis capabilities from both headquarters and district personnel to collect their ideas and impressions for new tools or improvement to existing tools. With that objective Task A documented PennDOT's current safety analysis tools, procedures and datasets and collected the expectations and goals of the Bureau of Highway Safety and Traffic Engineering (BHSTE) staff for the future of safety analysis at PennDOT.

PennDOT's Strategic Safety Goals and Objectives were developed as part of the CSHSIP that defines specific safety focus areas and sets goals for reduction of highway fatalities. The goals of PennDOT's CSHSIP provide the basis for most of the safety analysis expectations discovered during the PennDOT interviews of this task. Another primary expectation, relevant to the discovery of spatial tools, was that PennDOT's investment in CDART would be leveraged and not simply augmented and/or duplicated by a new tool.

In addition to expectations, the PennDOT interviews yielded a range of other common safety themes that contribute to a comprehensive safety analysis process. These themes included:

- **Dissect/rank/prioritize known clusters:** Closer examination of clusters to discover why crashes of a given type occurred with such frequency. Need to discern to what degree the conditions at that location make it a candidate for applying certain countermeasures.
- **What factors matter in prioritizing clusters?:** Need a way to looking for new ways to interrogate their databases to extract the most significant factors in crash causation and location ranking.
- **Pattern recognition:** Would like software that proactively scan the database and look for patterns that reveal the nature of crash history as it relates to road configuration, driver population, highway usage, and any number of other factors.

- **Combine linear and intersection clusters:** A clustering algorithm should be able to analyze crashes that are spatially related, regardless of roadway network configurations.
- **Include safety engineering and "soft side":** Future safety analysis methodologies and tools need to support all types of countermeasure programs not just engineering.
- **Integrate non-transportation data:** Future methodologies should include other environmental factors such as land use and demographics.
- **Integrate prescribed countermeasures:** Future analysis tools should include the concept of tracking safety engineering and soft-side countermeasures applied to roadway locations where crash clusters have been previously identified.
- **Utilize cost/benefit data:** Evaluate the cost of a proposed countermeasure versus the potential societal benefit of the expected reduction in crashes within a safety analysis system environment.
- **Integrate performance metrics:** Inclusion of data that measures the effectiveness of countermeasures would complete the safety management system cycle. This would provide management with useful information for high-level planning, and it would provide critical input to advanced predictive modeling tools.
- **Analyze cluster data for systemic improvements:** When countermeasures are shown to be effective in crash reduction, PennDOT wants to identify other locations where there is the likelihood that the same countermeasure will have the same positive effect.

The ultimate goal of these common themes is to improve the department's ability to proactively address highway safety issues rather than deal with them in a reactive manner. Additional details of PennDOT's specific safety goals, initiatives, expectations and existing crash analysis resources can be found in the Task A report.

2.4 Task B Analysis

Literature review and state/federal interviews, guided by PennDOT's areas of interest, found that there were no "commercial off-the-shelf" spatial tools for safety analysis, that FHWA's Safety Analyst application does not yet contain significant spatial capabilities, and that other state DOT's efforts toward spatial tools do not exceed PennDOT's current capabilities. Based upon best practice findings, key themes for specific low cost safety improvements and new alternative analysis methodologies were presented.

Task B compiled literature reviews, Web searches, and DOT and Federal Highway Administration (FHWA) interviews to summarize the current best practices in crash analysis and safety improvement. The literature review examined over 80 sources, and included analysis of documents generated by the Transportation Research Board (TRB), the American Automobile Association (AAA), and a state survey done by the Arizona DOT. Research also included a review of the FHWA's pooled fund project to develop

Safety Analyst. Safety Analyst is being built with four modules, Network Screening to identify sites for safety improvement, Diagnosis and Countermeasure Selection to design ways to address safety concerns, Economic Appraisal and Priority Ranking to establish priorities based on benefit:cost analysis, and Evaluation to analyze the benefits of safety improvements. Safety Analyst is, however, only in the beginning stages of adding geospatial capabilities and focuses primarily on the traditional engineering side of safety analysis without significantly addressing the soft side aspects of highway safety.

Interviews were conducted with FHWA, Iowa DOT, New York State DOT, Ohio DOT, and Washington State DOT. The research as well as the FHWA and state interviews focused on specific areas of interest. These areas of interest and the key findings are listed below:

- **Recognizing data patterns in crash records populations:** Most states are using traditional methods to identify patterns of high crash locations using information such as crash frequency, severity and trends. The indication from the research and interviews is that empirical Bayes techniques are the direction states are looking for future analysis. PennDOT currently uses traditional techniques but could benefit from more advanced statistical analysis tools.
- **Visualization techniques:** Task B discovered unique methods for symbolizing crashes on a map and the use of collision diagrams. PennDOT has the capability to use most of these visualization tools and currently generates collision diagrams manually.
- **Integrated data systems:** Including information outside of traditional crash analysis attributes such as locations of liquor establishments, schools, trauma costs, and climate zones was found to be of great interest, but not yet widely used. PennDOT has also integrated disparate data systems but on an ad-hoc basis only.
- **Integration of standard countermeasures and performance measurement:** Most states are comparing data from before and after countermeasure application and Safety Analyst will incorporate countermeasure effectiveness information when it is complete. The current processes used by PennDOT and other states for this analysis is highly labor intensive and is not done routinely.
- **Comprehensive approaches to highway safety analysis:** Isolated examples exist for DOT's using comprehensive analysis of the 4E's (Engineering, Enforcement, Education and Emergency Response) when determining appropriate combined countermeasures. Safety Analyst should be capable of this type of analysis but it is not commonly performed within state DOT's today.
- **Innovative highway safety data analysis techniques:** Our literature review found some interesting concepts in data mining. For the most part, applications are in health and law enforcement, not in transportation. Many DOT's have specific innovative programs for improved safety analysis including PennDOT, however the research and interviews found no applicable commercial software to address this need.

- **System-wide safety impact displays:** Identifying locations to apply specific countermeasures system-wide is not a standard procedure for any of the interviewed agencies and no COTS tools exist for this approach.
- **End-user capabilities:** Most DOT's, including PennDOT, have pushed geospatial technology tools down to the end user in their districts. These were found to be predominantly customized COTS products but web based applications like PennDOT's CDART are only beginning to be developed in a few states.

Overall, the key Task B findings showed that while there are unique programs and research being conducted by other states and research institutions the review of the current state of the practice shows that with CDART, PennDOT is already ahead of the field in the use of geospatial data and technology for highway safety. The research also did not discover any applicable Commercial off the Shelf (COTS) software packages that would fulfill PennDOT's wish for improved spatial analysis tools. In addition; while the FHWA's Safety Analyst application research shows the tool to hold great promise for analytical and performance based modeling, it remains perhaps a few years away from incorporating any viable spatial or soft side analysis capabilities.

Analysis of the Task B research findings did, however, solidify several key themes. One, overarching, theme (a common goal shared by each of the interviewed states) is to implement the most cost effective methodologies by focusing on low-cost improvements and leveraging existing resources to yield the greatest safety improvements. The research based recommendations, presented in the Task B report, for low cost improvements included:

- **Improved map visualization techniques:** Use existing GIS staff at PennDOT to improve the map presentation of data through advanced symbology to show crash aggregation as well as charts and graphs for enhanced map interpretation.
- **Organize regional homogeneous categories:** Develop additional homogeneous road type categories that account for regional differences such as weather and terrain.
- **Map the crash database to the Model Minimum Inventory of Roadway Elements (MMIRE) data model:** this effort would allow PennDOT to prepare for the release of FHWA's Safety Analyst software by identifying the gaps between the Crash Reporting System (CRS) and the Safety Analyst data model which is based upon MMIRE.

In addition to these low cost improvement recommendations there were four new methodology alternatives that stood out in the research. Determination of these methodology alternatives was guided by a set of overarching principals including: emphasis on the PennDOT areas of interest documented in Task A, recognizing the desire to implement spatial based tools, accounting for the ability to analyze and address soft-side safety management, and avoidance of any effort that would duplicate the functions or methods to be provided by Safety Analyst. With these principals in

mind, and with analysis of the state of the practice researched done in Task B, the following 4 methodology alternatives were recommended.

- **Spatial clustering:** Current PennDOT crash clustering methodology examines individual routes separately. Spatial clustering would provide the ability to identify crash clusters that occur on multiple routes such as at intersections and interchanges where route numbers change. These spatial clusters could uncover hotspots than the current clustering methodology has missed.
- **Data integration improvements:** Implement the ability to import and overlay other spatial data sets with the crash and roadway data. The contextual/spatial relationships with these new features, such as liquor licensed establishments or demographic data, could be used for alcohol related studies or defining new homogeneous road categories.
- **Countermeasure performance evaluation:** Develop a limited prototype to track performance metrics on countermeasure effectiveness that would provide support for PennDOT's Highway Safety Improvement Program.
- **Highway safety data relationships knowledge base:** Develop an initial framework for identifying and managing relationships between all potential highway safety analysis datasets. This prototype would analyze the logical, explicit or spatial data relationships between crash groups, causative factors and countermeasures.

These key research based findings and recommendations provided the foundation for the Task C Methodology Definition effort.

2.5 Task C Analysis

The methodology definition task presented 6 unique candidates for a proof of concept (POC). The knowledge base POC was selected for its long term value and broad applicability to the areas of interest, key research findings, and current PennDOT safety analysis process steps.

The objective of Task C was to compile and synthesize the information and guidance gathered in Tasks A and B to develop a new methodology for analyzing crash and roadway data and to identify the best candidate highway locations for safety improvements. An initial component of this task included the creation of a safety analysis methodology diagram depicting current PennDOT safety analysis processes. Documentation for the current methodology included the details of each process step and acknowledged the relevant and supporting Task B research findings for each. With the potential methodologies overlaid (Figure 1 – Task C Proposed Crash Analysis Tools), the diagram also provided a contextual framework for identifying which of the PennDOT safety analysis process steps might be extended/improved by a new analysis methodology.

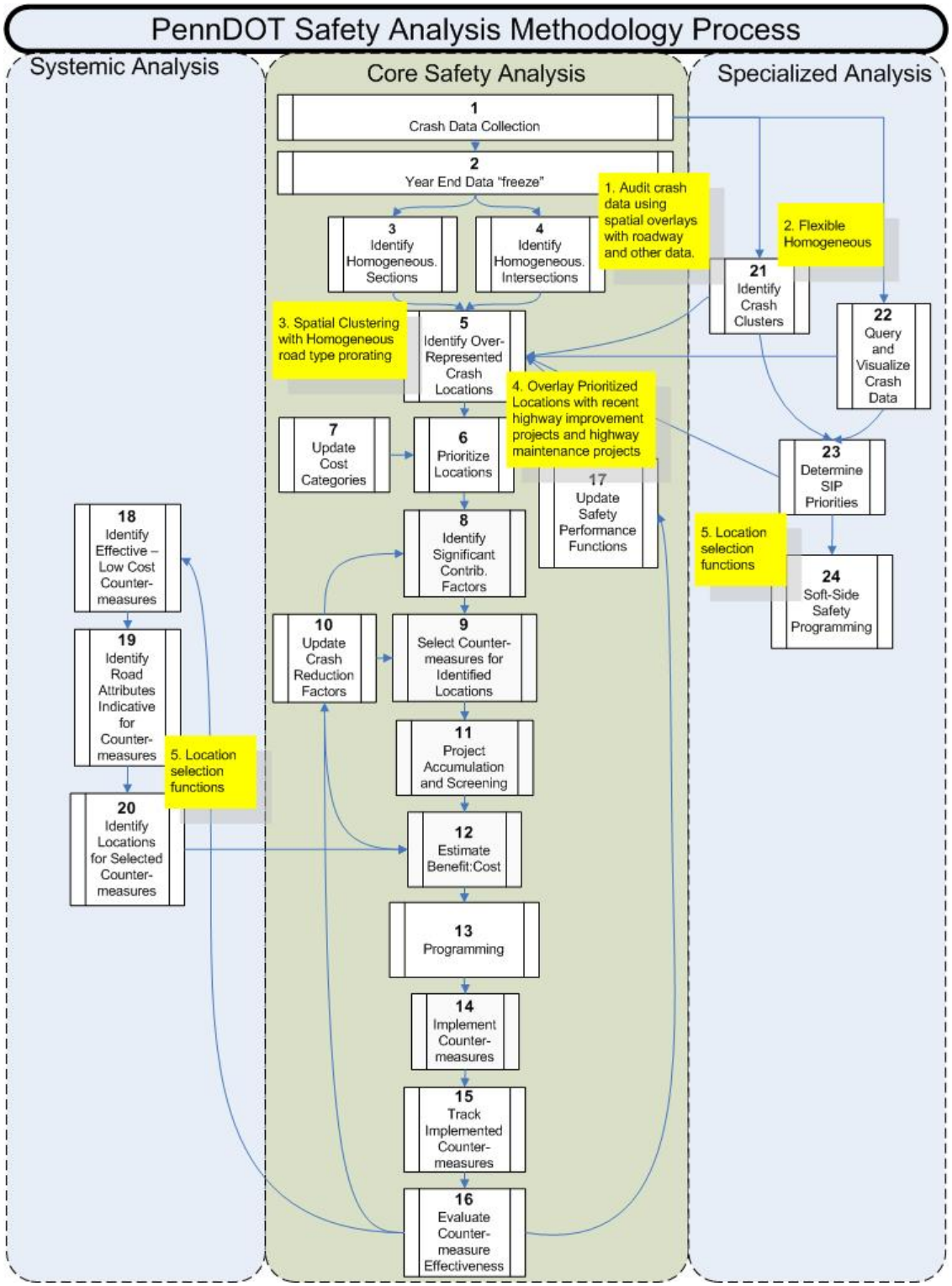
Based upon the PennDOT expectations and research findings six different proof of concept (POC) candidates were identified that could improve individual or multiple steps in PennDOT's current safety analysis process. These candidates are described below and Figure 1 shows areas in the process were these potential tools were identified.

1) **Audit Crash Location Data:** This audit process would cross check the police reported location, along with other data items on the crash report, against various roadway data and compute a confidence rating for the location of each crash. This audit would and provide a performance metric for crash location data quality thus improving the accuracy of the location priority process. While improving the critical data component for safety analysis this candidate does not directly address any of the PennDOT project expectations or research focus areas.
2) **Flexible Homogeneous Road Type Selection:** This POC would enable users to define custom road groupings based on specific attributes, including spatial datasets that are not currently used such as climate regions, land use, and proximity to features of interest such as bars and schools. This tool would leverage CDART by providing comparative data and addresses several research areas of interest.
3) **Spatial Clustering with Homogeneous Road Type Prorating:** PennDOT's current spatial cluster analysis does not detect clusters that span multiple adjoining routes/ramps. Using spatial, not route, based clusters would relate crashes that occur across multiple proximal routes and would provide the foundation for replacing current methods.
4) **Project/Maintenance Overlay:** Knowledge of recent/current highway improvement projects, maintenance or countermeasure implementations would allow for more accurate cluster prioritization. This tool would overlay this information with the cluster location priority list (LPL).
5) **Location Selection:** Inclusion of external data sources, such as location of bars or schools, within CDART for visual and spatial proximity analysis would promote the goal of integrating multiple data systems. This type of work is currently be handled on a case by case basis by the Geographic Information Division.

Highway Safety Data Relationship Knowledge Base: This candidate POC would provide a framework for identifying and managing relationships between all safety analysis data sets. The tool also has potential application across multiple steps in PennDOT's Safety Analysis Methodology.

The following sections provide an overview of the methodology definition task and are followed by a review of the chosen proof of concept results and implications for future use.

Figure 1: Task C Proposed Crash Analysis Tools



3 Proof of Concept Defined

3.1 Decision History for the Knowledge Base Proof of Concept

3.1.1 Presentation of Proof of Concept Options

At the conclusion of Task C (Methodology Definition) the content of the task report, including the PennDOT Safety Analysis Process diagram and analysis of six separate candidate proof of concepts (defined in Section 2.5 of this report), was presented to PennDOT. GeoDecisions presented an assessment of the six alternative methodologies using the following criteria, to facilitate PennDOT's selection of one of the methodologies to be developed as a proof of concept.

- **Process Diagram Reference:** Each step in the Safety Analysis Process where the POC could be used to assist that analysis was listed.
- **Methodology Description:** A brief description of the tool/method was presented.
- **Value Proposition:** A summary of the value the tool could add PennDOT safety analysis either qualitative or quantitative.
- **Proof of Concept Potential:** Specific example(s) of how the tool could benefit PennDOT's safety analysis capabilities.
- **POC % of Tool Development:** An estimate was provided noting what percentage of a final tool would be in place after implementing the POC.
- **Pros:** The potential benefits envisioned from implementing each candidate POC.
- **Cons:** The potential limitations/drawbacks envisioned from implementing each candidate POC.
- **Relevant Areas of Interest for Research:** Each of the important research areas identified by PennDOT that could benefit from implementation of the candidate POC.

After having been presented with the proof of concept options the PennDOT Project Panel decided on the Highway Safety Data Relationship Knowledge Base. This decision was made because the Knowledge Base was felt to have the most potential for long-term value due to its support for a greater number of the *Relevant Areas of Interest for Research* identified in Task B, and a more broad applicability to the current safety analysis methodology process steps defined in Task C. PennDOT expected this research project to find new technologies for safety analysis, and produce an innovative POC that would break new ground within the areas of interest for research while extending the current PennDOT safety analysis methodology. The Knowledge Base POC would move PennDOT to a level above the traditional safety analysis methodologies further emphasizing the research focus of the project.

> *The Knowledge Base POC can be directly applied to six of the eight areas of research investigated in Task B. These are listed in the Task C report. The safety analysis process steps where the Knowledge Base is applicable are discussed in section 4.2.1 of this report.*

3.1.2 Other Decision Considerations

PennDOT has expressed the desire to "take safety to the next level" by pioneering unique and modern analysis methodologies. Most traditional highway safety data analyses involve studying correlations among multiple data sets. The Knowledge Base was conceived as a repository of aggregated data combined with business knowledge that would serve as an information resource for a broad range of highway safety-related studies.

A Knowledge Base places virtually no limit on the amount of relevant data that PennDOT can introduce into this analysis. This flexible and more comprehensive approach will allow PennDOT to apply the tool to a broader range of research areas.

Also important to the decision to proceed with the Knowledge Base POC was the fact that CDART's mapping and query functionality can be used to complement the analysis done by the Knowledge Base proof of concept. Likewise the Knowledge Base output could be imported to CDART for map based display and/or provide the basis for new queries that are currently not part of the standard CDART analysis queries. The ability to capitalize on the existing CDART application and extend its analysis capabilities, without requiring redundant functionality be built, was one of the initial project goals and provided an additional basis for the decision to pursue the Knowledge Base POC.

PennDOT's decision to select the Highway Safety Data Relationships Knowledge Base as the methodology to pursue through proof of concept was made after careful consideration of all six options presented by GeoDecisions. The Spatial Clustering option was viewed as an attractive candidate, because it has the potential to identify high crash locations that may be obscured in other network screening methods. However, it was determined that the potential to identify additional crash clusters would not be of significant value to the Department, given current clustering tools already available. The other candidates (Audit Crash Location Data, Flexible Homogeneous Road Type Selection, Project/Maintenance Overlay, and Location Selection) were considered to be tools that could be addressed by current PennDOT resources and were not significantly differentiated from current methodology.

3.2 Overview of the Knowledge Base Proof of Concept

The knowledge base concept is an innovative methodology in highway safety analysis, as it incorporates concepts found in data warehouses and expert systems. It is capable of expansion to include virtually any data that can be correlated to crash or roadway data. The knowledge base provides a framework for identifying and managing relationships within data used for highway safety analyses.

In order to create a more comprehensive final report, the following generic Knowledge Base overview has been reprinted from the Task C Proof of Concept Methodology Report.

3.2.1 What is a Knowledge Base?

A Knowledge Base essentially acts as a domain information resource. While it has potential across an enterprise, its scope is seldom truly enterprise-wide. By its nature, it works within a more vertical, domain-based portion of the enterprise. A knowledge base provides more robust, contextual, searching capabilities. Searching within multiple data sets is made more efficient because the knowledge base has the ability to aggregate raw data into groupings that are specifically meaningful to domain areas within an organization.

The major component of a knowledge base is the capability for capture and storage of institutional business knowledge. Value can be added to raw data by linking stored business knowledge with that data. The business knowledge is an accumulation of known data behavior rules within a domain. Data behavior includes relationship rules within, and external to, specific data sets, data constraints based on real world scenarios, and how data should be interpreted as it relates to organizational policy. Business rules should include as much information as available that influence how data is used or interpreted. These rules are linked to specific data elements representing an example of how a knowledge base can transform raw data into valuable information.

Concepts related to knowledge bases are used in many other disciplines and include Business Intelligence, Decision Support System, Data Mining, Expert System, and Enterprise Information Systems.

3.2.2 What are the benefits of a Knowledge Base?

In general, the benefits of implementing a knowledge base application revolve around improved interpretation of data by managing the complex relationships that can be found among the glut of raw data. A knowledge base can identify what factors are statistically significant by applying statistical algorithms, contextually by looking for relationships between data elements, spatially by incorporating data attributes such as proximity and connectivity, and practically by constraining the analysis by known business rules. This improved data interpretation can be applied in various highway safety analyses involving crash causes, injuries, engineering countermeasures, and/or soft-side programs. The following are examples of improved highway safety data interpretation resulting from the application of a knowledge base.

- Analysis of disparate crash data to find significant factors in the potential causes of a given population of crashes.
- Ability to find significant correlations between minute details in diverse data sets such as crash, roadway configurations, countermeasures, etc.
- Using previously under-utilized data sets such as demographics, traffic citations, and environmental conditions to find significant correlations between highway safety factors/problems.
- Manage completely new and unique types of relationships involving business knowledge between crash data and external (off-site) data, unstructured data such

as documentation, financial data that drives funding assumptions or constraints, and spatial data providing context for analysis.

- Show how crash behaviors relate to or conflict with outside knowledge.

- More specifically, a knowledge base can benefit specific steps of the PennDOT Safety Analysis Methodology Process. See section 4.2.2 for details.

3.2.3 How is a Knowledge Base Used?

There are several potential components to a knowledge base application, but the required core data analyzing engine works in background of the application. This knowledge base "engine" would apply statistical algorithms, spatial queries, business rule constraints, and updates to soft business rules to prepare the data for further user analysis. Further analysis of the data can then be accomplished with several potential tools.

A **Relationship Viewer Tool** would allow a user to research and find information in ways that were previously time-consuming and tedious. By selecting any combination of data elements, a user could view statistically significant correlations within the constraints of known business rules. The user could then drill down into these correlations to follow the relationship links that are created by the background data processing.

A **Query Tool** would provide users the ability to find details on any subset, or intersected subsets of data aggregated by the core processing engine. Query selections could then have statistical functions applied to further refine and define the results.

A **Map Viewer Too**l could plot data generated by the relationship viewer tool and add additional spatial data layers providing richer context within which data could be analyzed. Standard GIS tools, such as intersection and buffer, could then be used to analyze the data. A use case for the map viewer tool would be to query accidents involving commercial vehicles, determine the data element relationships, and then plot those resulting crashes and relationships on a map that would include demographic and industrial/business park locations.

A **Report Generator** would plot standard reports based on the output of the Relationship and Query tools. Basic formatting could be made available for ad hoc report generation and a spreadsheet export tool would provide additional flexibility.

An **Administration Tool** is a necessary component of any knowledge base so that select users would have the ability to add, change, or delete business rules and related data groupings.

3.2.4 Knowledge Base Specifications

The prototype tool chosen for the PennDOT Spatial Tools project was crash analysis knowledge base since most highway safety data analyses involve studying correlations among multiple data sets. A knowledge base methodology requires a framework for identifying and managing relationships between all data sets involved with highway safety analyses. Essentially, the system would analyze the data to identify groups of crash locations, causative factors and countermeasures that are related – logically, explicitly, or spatially. The relationships would be identified and stored. The underlying code would reevaluate the relationships as new data is added. This would be an innovative methodology in highway safety analysis, as it would incorporate concepts found in data warehouses and expert systems. It would be capable of expansion to include virtually any data that can be correlated to crash or roadway data.

The basic data elements in the system are:

Incidents and clusters – Individual crashes and clusters identified by the current cluster analysis methodology. The crash data includes location data and by relation, road segment data where the crash occurred.

Causative factors – The causative factors are defined in the most generic way possible to take into account a wide variety of factors that could include environmental and other proximal data relevant to the crash.

Countermeasures – The Countermeasures include the all of the countermeasures identified in the current CDART tables.

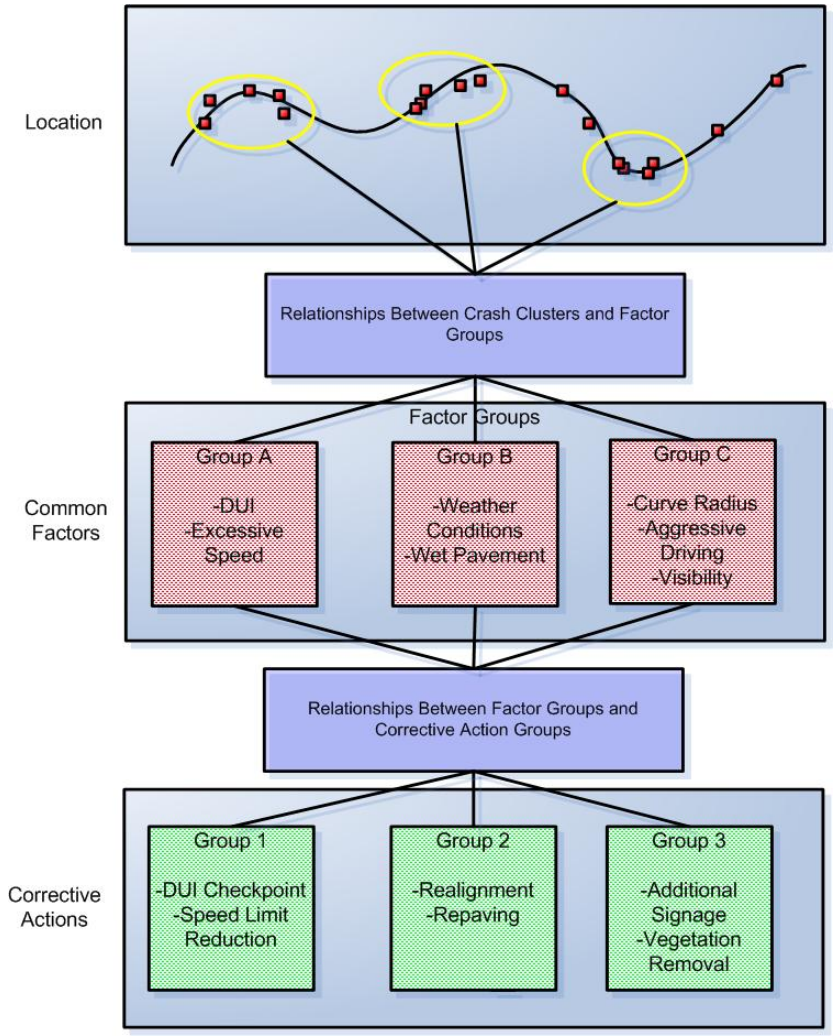


Figure 2: Proof of Concept Overview

The prototype analyzes the data to identify groups of incidents, causative factors, and corrective actions that are related. Relationships between these groups were identified and stored. The underlying code reevaluates these relationships as new data is added. See Figure 2 for an overview of the POC.

The prototype has several components. These components will contain the data, logical, and display elements. Each is briefly defined below and illustrated in Figure 3.

Raw Data – These are the raw data tables for crashes, factors, and corrective actions.

Business Rules and Relationships – This component contains the algorithms that will analyze the data as well as the Hard Business Rules (pre-defined rules) and Soft Business Rules (rules identified through multiple runs of the algorithm).

Aggregate Data Layer – This contains the initial groupings of the raw data elements.

Relationships Management Layer – This layer contains the end result of the analysis which is the set of relationships between individual and grouped data elements.

Relationship Viewer – This layer contains multiple user interfaces to view the data generated by the tool, including a map viewer, report viewer, query tool, and relationship diagram viewer.

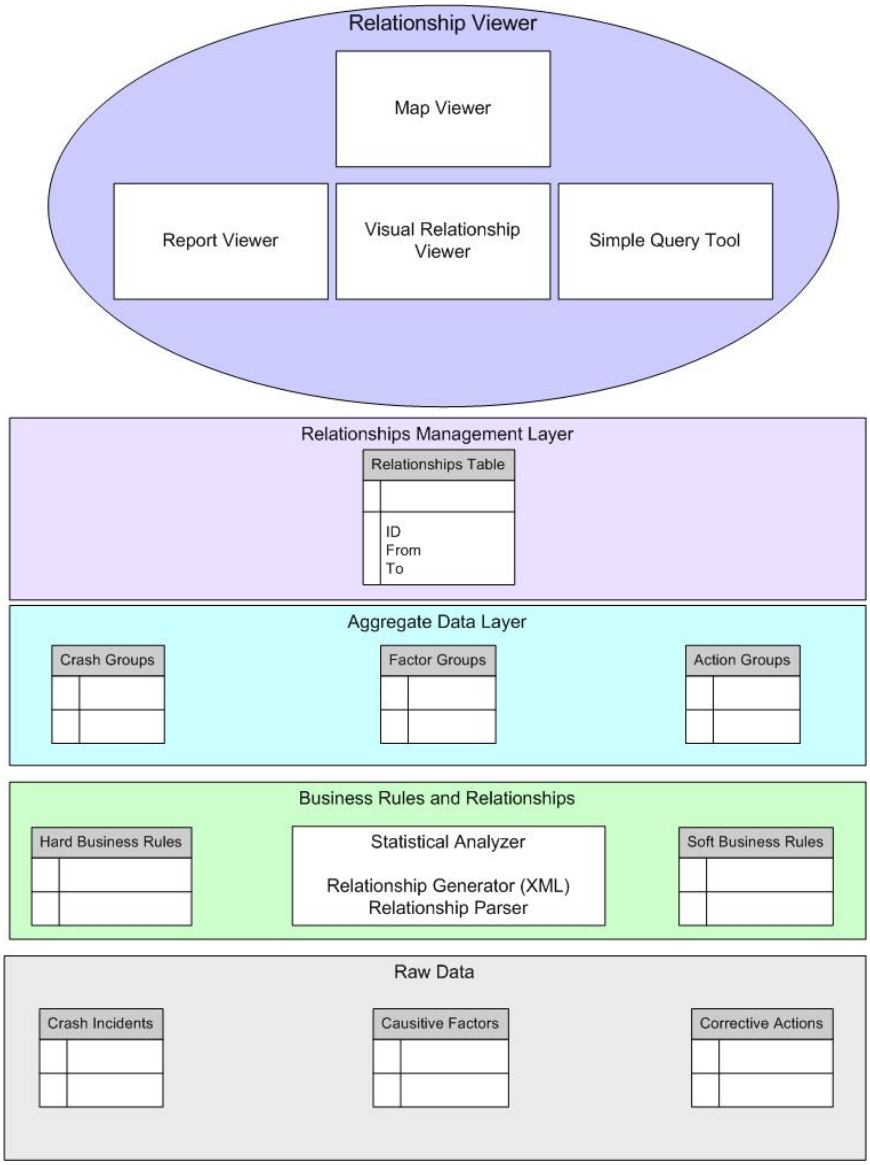


Figure 3: Prototype Component Structure

The operational workflow of the tool will require two passes of analysis – one to identify the generic statistical relationships and another to filter those relationships based on the hard and soft business rules. Figure 4 illustrates the workflow of the tool.

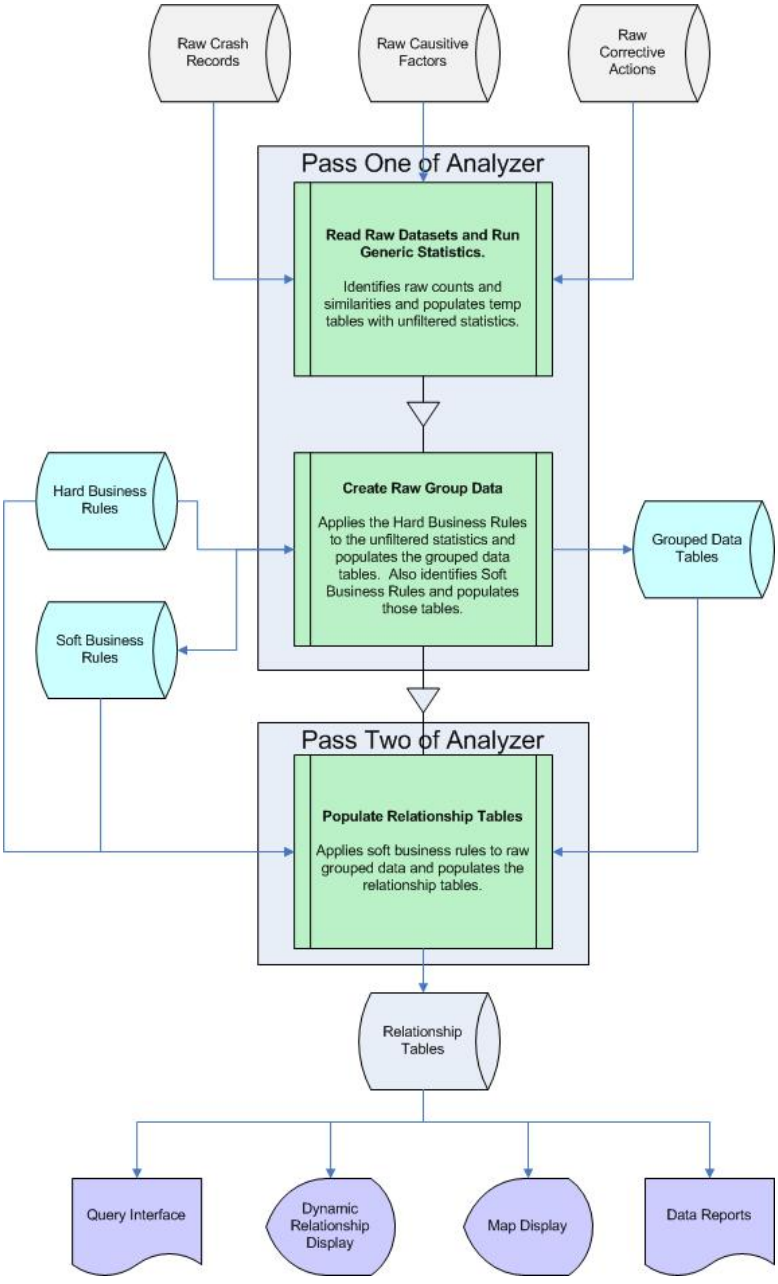


Figure 4: Prototype Workflow

3.2.5 Proof of Concept Requirements and Design

Detailed requirements and design documentation were produced during development of the Knowledge Base proof of concept. This documentation can be found in the Task D: Proof of Concept report.

4 Proof of Concept Results and Implications

4.1 Analysis of the Proof of Concept Results

The proof of concept demonstrated the feasibility of constructing a knowledge base that can provide PennDOT

- A new way of viewing crash and roadway data to reveal patterns of significant relationships;
- Integration of safety engineering and safety programming in the analysis of countermeasures
- Support for multiple approaches to highway safety analysis, to include remediating high crash locations, diagnosing safety focus areas, and system-wide countermeasure study.

The Knowledge Base application developed for this research project, while still a Proof of Concept, contains many of the key functions described at the conceptual level in the Task C Methodology Report. The testing of this POC application was able to successfully demonstrate the abilities of the application thus validating it as proof that the Knowledge Base concept worked as intended. This section provides an analysis of the test cases used and the results established through implementation and testing using PennDOT crash data.

The Proof of Concept Design (Task D Report) describes in detail the scenarios and use cases used to demonstrate the processes and functions of the Knowledge Base application. At PennDOT's suggestion, scenarios focused on analyzing "run off the road" crashes were input into the knowledge base analysis engine. These scenarios (13 in all) included the combinations of roadway factors, driver factors and all factors for identification of significant correlative relationships. The use cases covered three approaches for analyzing highway safety, focus areas, hot spots, and system wide countermeasures as described below.

- **Focus Areas:** This use case was derived from PennDOT's 2006 Comprehensive Strategic Highway Safety Improvement Program which calls out vital safety focus areas such as aggressive and impaired driving. The POC testing for Focus Areas was conducted by selecting "Drinking Driver" as the causative factor and analyzing all factor groups where "Drinking Driver" is included. The crash groups that contain any of the factors in these factor groups were then reviewed for potential Corrective Action Groups. This approach unveiled a strong correlation between "Drinking Driver" and "Unbelted" and presented the relationship to two corrective actions.
- **Hot Spots:** This type of analysis aimed to diagnose a problem at a specific location and determine the best means to address the safety concerns. This use case filtered the crashes for analysis by a specific county and route to produce crash, factor and action relationships associated with just the records included within the spatial constraints. The use case was selected in part because it

conforms to both PennDOT's and federal Highway Safety Improvement Program guidelines. This application was able to show comparisons in number of related crash groups correlated to a chosen corrective action and map the individual members of the crash group.

- **System Wide Countermeasures:** This use case was aimed at identifying candidate locations, across the entire highway system, for low cost safety engineering improvements. The POC testing looked at edge rumble strips as a system wide countermeasure strategy by first selecting it from the Corrective Action list. All Corrective Action groups with that action are then populated in the interface allowing the user to interactively examine the related factor groups searching for groups whose profile is conducive to and appropriate for the installation of edge rumble strips.

The results of these Use Cases are more thoroughly discussed below by demonstrating how each of the key Knowledge Base functions performed in the proof of concept application.

- **Statistical Analysis Engine:** This automated function recognizes data patterns where crash record attributes are highly correlated. All of the above use cases utilized this analysis engine by successfully processing the input data. The statistical analysis engine was shown to run independently of user intervention and produce the crash groups that act as homogeneous crash categories, forming the basis of subsequent user driven analysis. Without any other knowledge base functions, the data produced by this analysis engine is useful for crash group prioritization and as baseline metrics for other safety applications/analysis.
- **Business Rules:** The use cases for this POC applied business rules to the knowledge base by defining scenarios relevant to highway departure crashes. Rather than allowing a virtually unlimited number of scenarios to be analyzed, trained users can apply their business knowledge to constrain the statistical analysis engine by segmenting the input data into more logical scenarios. In this case the scenarios grouped data into categories based on attributes for driver or roadway factors. The POC application provided the ability to apply these business rules, establishing a more homogeneous and highly correlated data set for the statistical analysis.
- **Determine and Maintain Relationships:** The POC application was shown to generate relationships between crash, causative factor and corrective action groups providing users the ability to initiate analysis by examining these relationships based on selections from any of the three groups. Each of the use cases made use of this functionality starting with crashes for "hot spot" analysis, causative factor groups for "focus areas" and corrective action groups for investigating possible "system wide countermeasures".
- **Summary Calculations:** The POC also demonstrated the ability and benefit of performing summary calculations on the relationships generated. The POC displayed a summary for the number of crash groups which, for example, would update based upon the selection of a specific countermeasures. Showing

variations in the total number of crash groups affected for each countermeasure alternative analyzed allows the user to more effectively perform evaluations.

- **Map Output:** Key to capitalizing on the spatial component of this research project was the ability to produce a map of the analysis results. The prototype application allows the user to produce a map of the crash records present in selected crash groups. This was demonstrated for the hot spot use case with the output shown in the task D report.

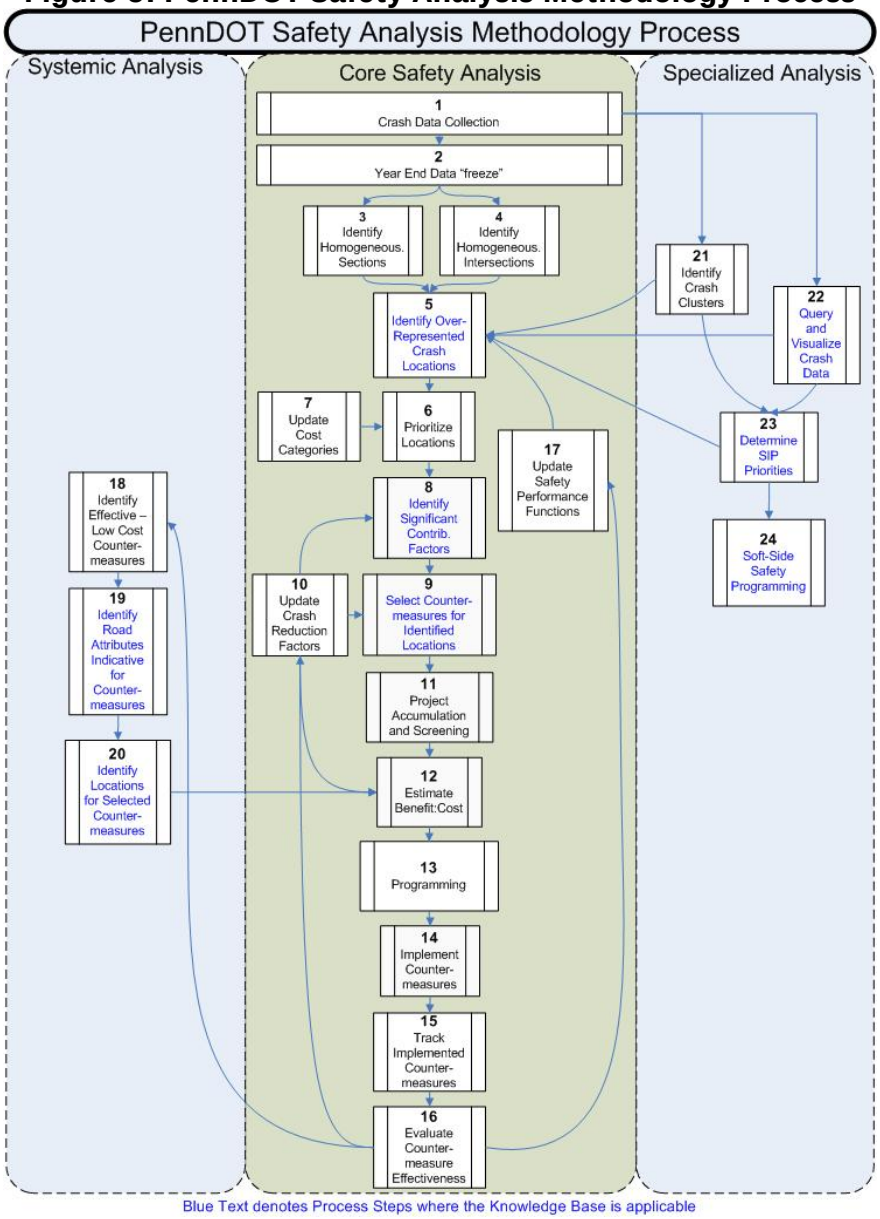
The results generated from the prototype application were evaluated against both practical use cases and the applications functional requirements. These results were shown to fulfill the intent of a proof of concept and yielded very positive implications for expanded use in the future. The final section of this report explores the potential for applying this research POC in specific and tangibly beneficial ways at PennDOT.

4.2 Implications of Results to PennDOT's Safety Analysis Process

Applications for the knowledge base exceed the use cases demonstrated in the proof of concept. It has the potential to *add value* to other highway safety tools, data, and methods in PennDOT, including CDART and the emerging crash location priority technique documented in PennDOT's Highway Safety Manual.

The following diagram depicts the three primary types of Safety Analysis done by PennDOT. Systemic Analysis includes the steps involved in determining system-wide safety measure implementations. Specialized Analysis covers the more ad-hoc processes, carried out partially by CDART, that serve more specific analysis functions. The Core Safety Analysis depicts PennDOT's Safety Improvement Program (SIP), the primary safety engineering process in the Department. The diagram denotes areas where the Knowledge Base Proof of Concept tool can be applied and these areas are further defined below.

Figure 5: PennDOT Safety Analysis Methodology Process



4.2.1 Process Areas where the Knowledge Base is Applicable

The following describe how the Knowledge Base is applicable to specific areas of PennDOT's Safety Analysis Methodology Process. The knowledge base does not replace this legacy methodology, rather it leverages and extends it with new data analysis capabilities. Each of the process steps below can benefit from the flexibility of the Knowledge Base analysis capabilities.

- Identify Over-Represented Crash Locations (5) – Allows the exploration of new relationships between crash clusters other data such as weather patterns,

emergency services, or grant programs which may indicate new locations for safety focus.

- Prioritize Locations (6) – Building new relationships to crashes in the knowledge base between unstructured data such as funding streams and hard/soft countermeasure programs can provide distinguishing criteria for improved prioritization.
- Identify Significant Contributing Factors (8) – By tracing through statistically correlated combinations of factors in crash, roadway, and environmental data, users can identify contributing factors that may not have previously been recognized.
- Select Countermeasures for Identified Locations (9) – Suggest countermeasures based on newly discovered data linkage combinations and business knowledge regarding the appropriate application of those countermeasures to better address safety at specific locations.
- Identify Effective Low Cost Countermeasures (18) – Find significant statistical relationships among crash rates, countermeasure types, financial constraints, and practical knowledge to help choose the most efficient and effective countermeasure(s).
- Identify Locations for Selected Countermeasures (20) – Pinpoint locations for specific countermeasures using newly discovered data relationships such as identifying that a second type of crash is reduced by a countermeasure previously reserved for other scenarios. This will enable ongoing updates to PennDOT's list of countermeasures and their potential applications.
- Query and Visualize Crash Data (22) – Using the results of statistical analysis and the relationships discovered in the data, users can identify "correlative clusters" where the distribution of crashes with significant attribute correlation are visible when mapped thus enhancing the existing functionality of CDART.
- Determine SIP priorities (23) – Lets users investigate and discover new opportunities for specific hard/soft-side programming combinations based on new statistical correlations found in the crash data.

4.2.2 Implications for Wider Implementation of the Knowledge Base

The Knowledge Base POC's success at delivering the technical scope of the project has been demonstrated through the results analysis presented in the Task D report and sections above. These initial results provide validation of the tool's usefulness to PennDOT's safety analysis process through direct applicability to process steps, while also advancing PennDOT's technical capabilities within its research areas of interest.

This project and the prototype Knowledge Base, however, does more than just satisfy the technical scope by delivering evidence that the concept of a data relationship knowledge base can be implemented in a useful manner at PennDOT. The prototype application provides both a tangible and theoretical foundation for advancing PennDOT's safety analysis capabilities and has substantial opportunities for extensibility. Evidence of this extensibility is shown by the inherent knowledge base ability to incorporate and analyze virtually limitless datasets for correlations and relationships. The addition of

new datasets will allow for incorporation of spatial and soft-side elements that can further refine the profile of any given crash. Extended attribute based profiles can uncover correlations with crash factors that were previously unknown. Creation and management of these correlative relationships add a potential geometric gain in PennDOT's analysis capabilities.

The implications of these conceptual knowledge base capabilities for PennDOT are supported by the description of several specific implementation ideas listed below. These potential implications were suggested directly by PennDOT as well as implied through the original Task A expectations.

4.2.2.1 Specific Implications for PennDOT

The Task D Proof of Concept examined several specific use cases showing ways that PennDOT can use the POC tool with their own data to investigate safety improvement program focus areas, diagnose hot spots and evaluate system wide countermeasures. These use cases met the requirement to prove the viability of the prototype application but merely scratched the surface of the potential knowledge base applications.

PennDOT has suggested that perhaps the most immediate implication for the broader use of the Knowledge Base at PennDOT is its potential integration with CDART. The Knowledge Base output of unique crash record, factor, and corrective action relationships could be imported to CDART for map based display. *The output of map-able data has already been demonstrated in the POC "Focus Area" use case.* The new and unique relationships generated by the Knowledge Base can also provide the basis for queries that are currently not accounted for in CDART. For example, CDART allows users to generate queries from a finite set of attributes and relies upon those users to formulate viable queries. Creating new CDART queries based upon output from the Knowledge Base makes best use of both tools by providing that CDART queries, and the analysis derived from them, are based on statistically valid relationships. The ability to capitalize on the existing CDART application and extend its analysis capabilities would satisfy one of the primary PennDOT goals, to make the best and most efficient use of existing tools and resources.

In addition to the integration of the Knowledge Base with CDART several other potential uses were discussed or realized once the POC results were analyzed within the context of the project findings. The Task A Data Review reported on a goal of the Safety Management Division of BHSTE to incorporate additional spatial and non-engineering, soft-side, data into the safety analysis process. PennDOT has performed GIS proximity analysis for special projects in the past but this type of spatial analysis has not been integrated as a routine part of the analysis methodology. Incorporating additional spatial based attributes in the knowledge base analysis, such as unique attribute values for crash records that designate a climate region or buffer distance threshold from alcohol establishments, includes these spatial based contributing factors when defining significant relationships. In the same manner attributes that may influence the use of soft-side countermeasures could be included in the Knowledge Base analysis engine.

This project has shown that a knowledge base is capable of generating new homogeneous crash categories based on the unique and diverse attribute relationships discovered by the statistical analysis engine. These new homogeneous crash categories are valuable for targeting specific countermeasures that may provide more effective remediation when used to address crashes with these newly discovered distinct attribute profiles. The Knowledge Base can be used to reveal opportunities to wield PennDOT's countermeasure arsenal in a new and more effective manner.

Each correlative cluster, those crashes with significant attribute correlation, generated by the knowledge base has a "cluster center" which is the statistical average crash for that crash group. By analyzing the members of that crash group that deviate from the average, PennDOT can identify crashes that are most in need of treatment. An analyst could map those crashes identified with higher than average values in order to visualize the locations where these above average crashes occur and subsequently focus improvements on those problem areas.

Upon viewing the results of the POC use case testing it was also evident to the project team and PennDOT panel that summarizing the results, especially when comparison of multiple analysis scenarios is needed, would add value to the tool and make better sense of its output. These summaries could include details such as average values for each crash group, sophisticated severity rankings based on the knowledge base's more diverse set of crash attributes, and countermeasure rankings based on benefit:cost calculations.

The knowledge base has been proven to be compatible with, complementary to, and even capable of extending, many of the current tools and methods used at PennDOT for highway safety analysis. Examples of these include CDART, conventional location prioritization methods/algorithms, and numerous analysis process steps illustrated by this project. The knowledge base also has potential to appreciably contribute to future highway safety analysis initiatives such as the imminent Safety Analyst software from FHWA and the newly emerging solutions being documented in PennDOT's Highway Safety Manual. For example: PennDOT is looking into methods where the entire roadway system and crash database are analyzed to identify where just a small percentage of the network hosts a disproportionately large percentage of total crashes. The knowledge base can be used to both investigate and validate these new methodologies through input of purposeful scenarios and business rules, hypothesis based crash, factor and countermeasure filtering and performing custom summary calculations on results.

The Knowledge Base's ability to expose new datasets, supply correlation analysis abilities, and integrate into both existing and future tools and methodologies will enable PennDOT to enhance and even transcend traditional analysis by uncovering new substantive data relationships vital to improving the safety of Pennsylvania roads. This groundbreaking technological ability is occurring nowhere else in the country and will certainly set PennDOT at the forefront of highway safety analysis and support the States objective to "take safety to the next level".