Pitney Bowes

Spatio-Temporal Networks:

Analyzing Change Across Time and Place

WHITE PAPER

By:
Jeremy Peters ,
Principal Consultant, Digital
Commerce Professional Services,
Pitney Bowes

ABSTRACT

ORGANIZATIONS ARE GENERATING POWERFUL INSIGHTS BY ANALYZING CHANGE IN SPATIO-TEMPORAL NETWORKS IN MANY APPLICATIONS, SUCH AS WEATHER RISK ANALYSIS. THE RAPID GROWTH IN SIZE, VARIETY AND UPDATE RATE OF SPATIO-TEMPORAL DATA IS CREATING NEW CHALLENGES AND OPPORTUNITIES TO EFFICIENTLY STORE, VALIDATE, PROCESS AND ANALYZE SPATIO-TEMPORAL NETWORKS WITH LARGE TIME-SERIES DATA. THIS PAPER DESCRIBES THE CHALLENGES AND TRENDS IN BIG DATA ANALYTICS FOR SPATIO-TEMPORAL NETWORKS. IN ADDITION, IT OUTLINES A PROOF-OF-CONCEPT USE CASE FOR HISTORICAL TORNADO EVENT RISK ANALYSIS THAT IS IMPLEMENTED USING THE PITNEY BOWES® SPECTRUM™ TECHNOLOGY PLATFORM SOFTWARE FOR CONFIGURING AND RUNNING BATCH JOB AND REAL-TIME WEB SERVICES INTEGRATED WITH R DENSITY BASED SPATIAL CLUSTERING AND SQL SERVER SPATIAL. LEARN HOW PITNEY BOWES SPECTRUM™ SPATIAL FOR BUSINESS INTELLIGENCE SOFTWARE CAN BE USED TO PROVIDE A WEB-BASED MAP VIEWER TO ANALYZE AND VISUALIZE TORNADO EVENT SPATIAL CLUSTERS OVER TIME, AS WELL AS THE MAGNITUDE OF AND PROPERTY DAMAGE CAUSED BY TORNADO EVENTS THROUGH SURFACE DENSITY MAPPING.

Challenges and trends: big bata analytics for spatio-temporal networks

Spatio-temporal networks encompass the spatial relationship among locations and the time dimension of the data that those locations represent. Analyzing change in spatio-temporal networks is important in many applications in commerce, transportation, electricity and gas distribution, telecommunication networks, air/water/land quality monitoring and weather risk analysis, among many others. The size, variety and update rate of spatio-temporal data is growing fast, and quality is often an issue, as it comes from databases, web application logs, industry specific transaction data and location aware devices like mobile phones and many kinds of sensors. Spatio-temporal data is often massive in size. It is not uncommon to see visualization systems of spatio-temporal data, such as map viewers, that focus on the spatial dimension and do not effectively address the temporal dimension. These new challenges need to be addressed to efficiently integrate, store, validate, process and analyze spatio-temporal networks with large time-series data.

Global data growth, in general, and spatio-temporal data growth, specifically, are characterized by variety, volume, velocity and veracity. A new complexity is caused by many new data types that are now being collected in addition to master and transactional data such as semi-structured data (e.g., email, electronic forms, etc.), unstructured data (e.g., text, images, video, social media, etc.), as well as sensor and machine-generated data. The volume of data that companies and governments are collecting is growing rapidly. They must now deal with data sources that are hundreds of terabytes or more that need to be stored and analyzed. The rate/velocity at which data needs to be created, processed and analyzed, often in real-time, is increasing in many applications, such as financial market and traffic data. The new variety, great volume and rate at which data needs to be processed and analyzed is creating even greater data veracity/data quality challenges. Data needs to be cleaned and data quality needs to be implemented before confidence can be established in the results of any analytics.

Big Data analytics typically involves analytical workloads where data variety, data volume and/or data velocity play a part. A variety of technology platforms is used for solution implementation, depending on the nature of the data and the analytical requirements. These technologies include relational and analytical DBMS (SQL Server and data warehouse appliances, respectively), non-relational data management platforms (e.g., Hadoop platform), NoSQL data store (e.g., graph database, such as Neo4j) and Stream processing software. Stream processing software, is used to support the automatic analysis of events as they happen in real-time or near real-time to identify significant patterns in data streams and trigger action to respond to them. In this case, analysis of the data, often using predictive or statistical models, usually takes place before the data is stored. A Rules engine can be used to automate decision-making and action taking.

LOCATION IS OFTEN USED TO RELATE AND JOIN DISPARATE DATA SOURCES THAT SHARE A SPATIAL RELATIONSHIP.

Enterprise information management software provides the basis for an extended analytical environment. One of the strengths of this software is its ability to define data quality and data integration transforms in graphical workflows, such as the Spectrum Technology Platform proof-of-concept data flow described below for historical tornado event risk analysis. Analytics can now be pushed down into analytical databases and into Hadoop, and analytics, rules, decisions and actions can now be added into information management workflows to create automated analytical processes. Workflows can be built and re-used regularly for common analytical processing of both structured and un-modelled, multi-structured data to speed up the rate at which organizations can consume, analyze and act on data. Workflows can be implemented as batch jobs and real-time web services.

Location intelligence often plays an important role in organizing and using the big data. Location is often used to relate and join disparate data sources that share a spatial relationship. Location intelligence visualization can help identify patterns and trends by seeing and analyzing data in a map view with spatial analysis tools such as thematic maps and spatial statistics. Location intelligence can help find "data needles in a data haystack" by using spatial relationships to filter relevant data.

Temporal location intelligence is being applied in many applications using spatio-temporal clusters, simulation and visualization; change detection; map animation and movement tracking. Temporal data is being integrated into location intelligence databases so that structured queries can executed against both spatial and temporal attributes can be executed. Providing temporal operators that can be used with spatial data operators is needed to effectively analyze spatio-temporal data. Location intelligence software, such as the Spectrum Platform, can provide data processing, visualization and analysis tools for both the time and geographic dimension of data that helps expose important insights and provide actionable information.

Use case: historical tornado event risk analysis POC

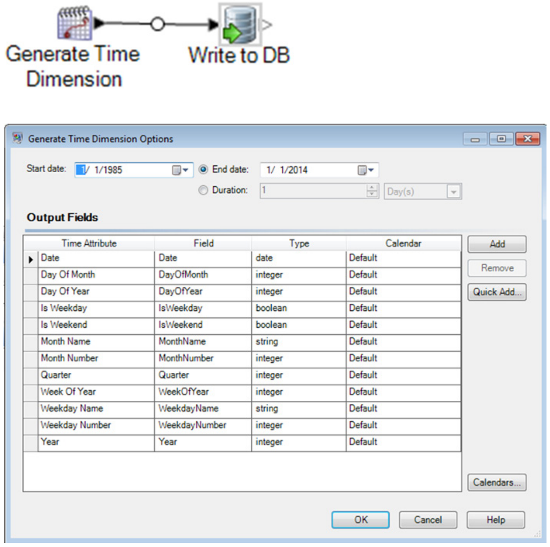
The Spectrum™ Technology Platform is service-oriented architecture (SOA) platform with a server that supports modules for a broad set of capabilities for data quality, master data management, analytics, location intelligence and data integration. The Spectrum server provides the ability to handle large data sets, combined with advanced clustering and in-memory caching, can process large volumes of data coming in at high velocity. The Spectrum Enterprise Designer client tool gives users drag-and-drop capabilities (e.g., Generate Time Dimension, Geocode U.S. Address, Get Travel Boundary, Query Spatial Data, etc.) onto a workflow designer to construct business-process data flows in the form of batch jobs or web services.

A proof-of-concept application for the analysis of historical tornado event data was implemented using the Spectrum Platform to visualize on a map tornado event spatial clusters in relation to their magnitude and property damage caused for a user defined study area (drive time or radius around a specified address) and for a user-defined time period (Jan 1, 2000 to Jan 1, 2010 and for the months of March, April and May/tornado season). Using the application, users can compare tornado event results for different time periods (tornado season vs. not tornado season) and/or different study areas.

The first implementation step involved adding the required time dimensions to the historical tornado events in order to prepare the historical tornado event records for analysis by time. Spectrum data flows were created using Spectrum's Generate Time Dimension and Record Joiner capabilities, to generate a time dimension table and add the required time dimensions to the tornado events in an SQL Server table. Spectrum's Time Dimension stage provides accurate time-based calculations on historic data without using complex SQL calculations. For example, time dimensions enable you to analyze your data by workdays versus holidays, weekdays versus weekends, by fiscal periods or in this case by tornado seasons.

A batch job and real-time web service data flows were configured and integrated with R Density based Spatial Clustering and SQL Server Spatial in order to:

- Select historical tornado events from SQL Server that took place within specified dates, within specified months (e.g., tornado season) and within a specified drive distance or radius around a specified address (e.g., study area). A POC is being developed where this type of spatial query can be run against data stored in Hadoop Distributed File System (HDFS) to provide high-performance access to large data sets.
- Calculate spatial cluster for each tornado event using R Density based DBScan Spatial Clustering. R is an open source software programming language and software environment commonly used for statistical computing and data analysis. Clustering algorithms can reveal distribution patterns on spatiotemporal data. It is possible to find regions and periods that have great Tornado density. Tornado events that rarely occur in given regions can also be detected.



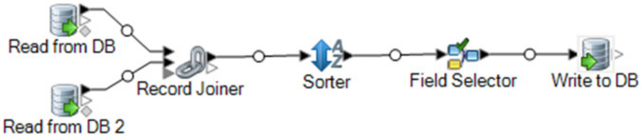
Start date: 1/ 1/1985 ● End date: 1/ 1/2014

○ Duration: 1 Day(s)

Output Fields

Time Attribute	Field	Type	Calendar
Date	Date	date	Default
Day Of Month	DayOfMonth	integer	Default
Day Of Year	DayOfYear	integer	Default
Is Weekday	IsWeekday	boolean	Default
Is Weekend	IsWeekend	boolean	Default
Month Name	MonthName	string	Default
Month Number	MonthNumber	integer	Default
Quarter	Quarter	integer	Default
Week Of Year	WeekOfYear	integer	Default
Weekday Name	WeekdayName	string	Default
Weekday Number	WeekdayNumber	integer	Default
Year	Year	integer	Default

The Generate Time Dimension dataflow and stage options create a table with the required time dimensions

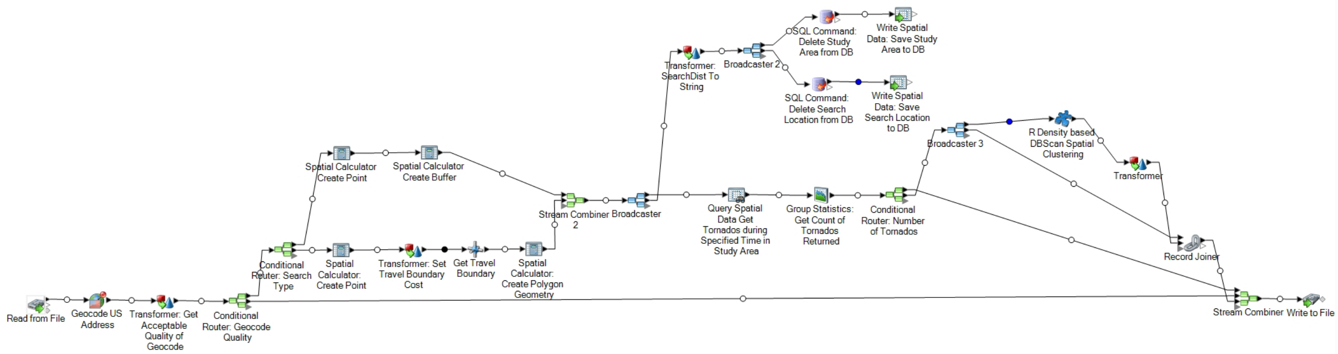


The Generate Tornado Event Time Dimension data flow adds the required time dimensions to the tornado event table

The Spectrum data flows were created to accept the following attributes as input for analysis.

- **StartDate & EndDate:** Time period of tornado events to analyze (e.g., 1985-01-01 to 2011-01-01)
- **Months:** Months during the year of tornado events to analyze, such as March, April, May, to represent tornado season
- **AddressLine1,City,StateProvince,PostalCode:** U.S. address that represents the starting point of the study area to analyze
- **SearchDistance:** Distance in miles that will be used to create a drive distance or radius around the specified address
- **SearchType:** Drive distance or radius around the specified address

LOCATION INTELLIGENCE CAN HELP FIND DATA "NEEDLES IN A DATA HAYSTACK" BY USING SPATIAL RELATIONSHIPS TO FILTER RELEVANT DATA.


Read from File
Geocode US Address
Transformer: Get Acceptable Quality of Geocode
Conditional Router: Geocode Quality
Conditional Router: Search Type
Spatial Calculator Create Point
Spatial Calculator Create Buffer
Spatial Calculator: Create Point
Transformer: Set Travel Boundary Cost
Get Travel Boundary
Spatial Calculator: Create Polygon Geometry
Stream Combiner 2
Broadcaster
Transformer: SearchDist To String
Broadcaster 2
SQL Command: Delete Study Area from DB
Write Spatial Data: Save Study Area to DB
SQL Command: Delete Search Location from DB
Write Spatial Data: Save Search Location to DB
Query Spatial Data Get Tornados during Specified Time in Study Area
Group Statistics: Get Count of Tornados Returned
Conditional Router: Number of Tornados
Broadcaster 3
R Density based DBScan Spatial Clustering
Transformer
Record Joiner
Stream Combiner
Write to File

The Historical Tornado Event Risk Analysis data flow finds tornado events with a specified study area for a specified time period and calculates the spatial cluster value for each tornado event

The data flow processing of the specified locations include:

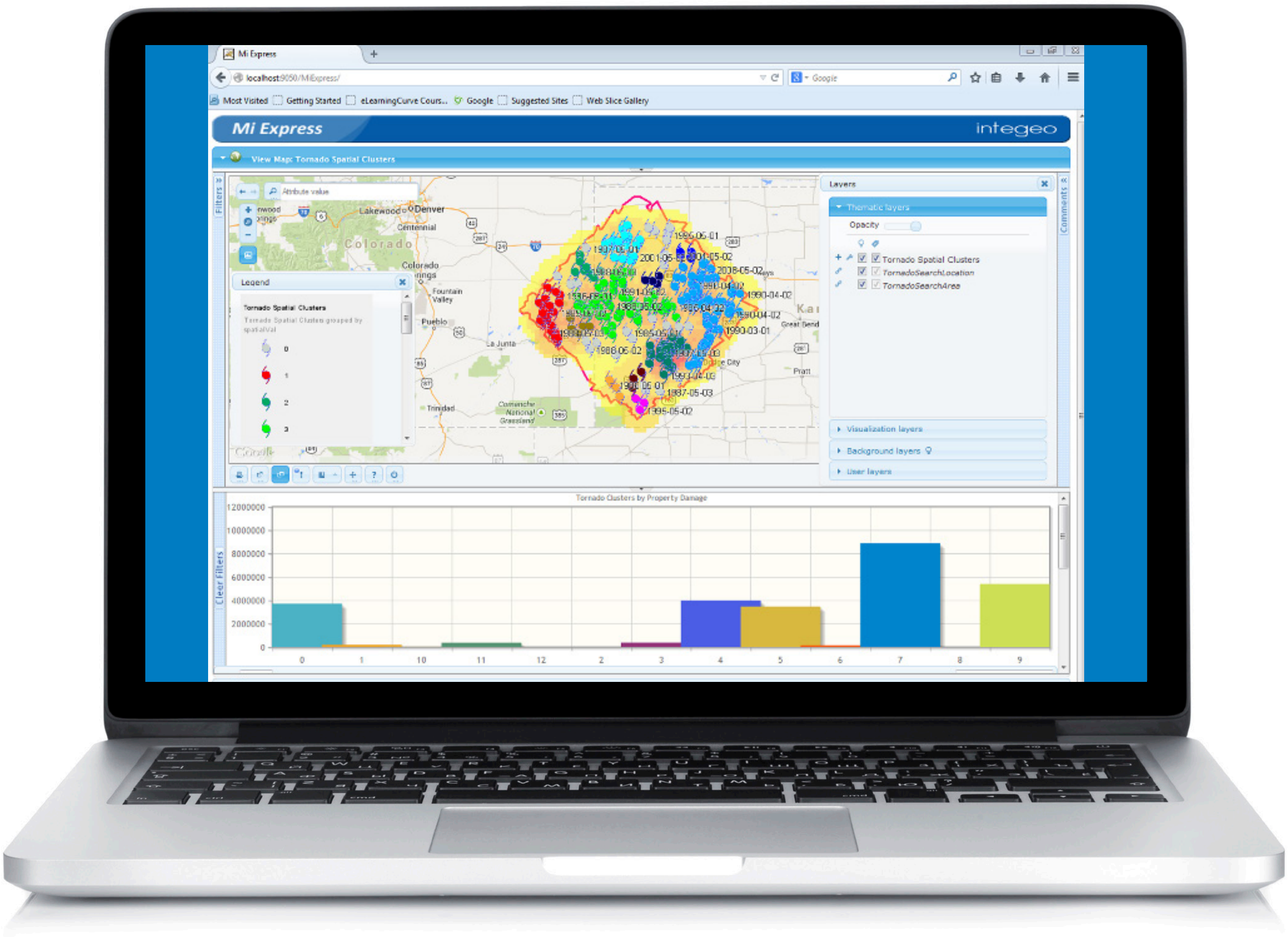
- Standardizing and geocoding the input address
- Generating a drive distance or radius boundary around the geocoded address using the specified search type and distance
- Selecting historical tornado events from SQL Server that took place within the specified dates, within the specified months and within the specified drive distance or radius around the specified address using temporal and spatial operators
- Calculating the spatial cluster for each tornado event using R Density based DBScan Spatial Clustering

For this application, Spectrum Spatial for Business Intelligence is used for web mapping visualization and analysis of the tornado event spatial clusters and attribute output. It adds mapping and spatial analytics into such business intelligence applications as Cognos, Business Objects, QlikView and MicroStrategy. It enables bi-directional analysis between data visualized on maps and other more traditional representations such as tables of data, charts and reports to help discover previously hidden information and data relationships.

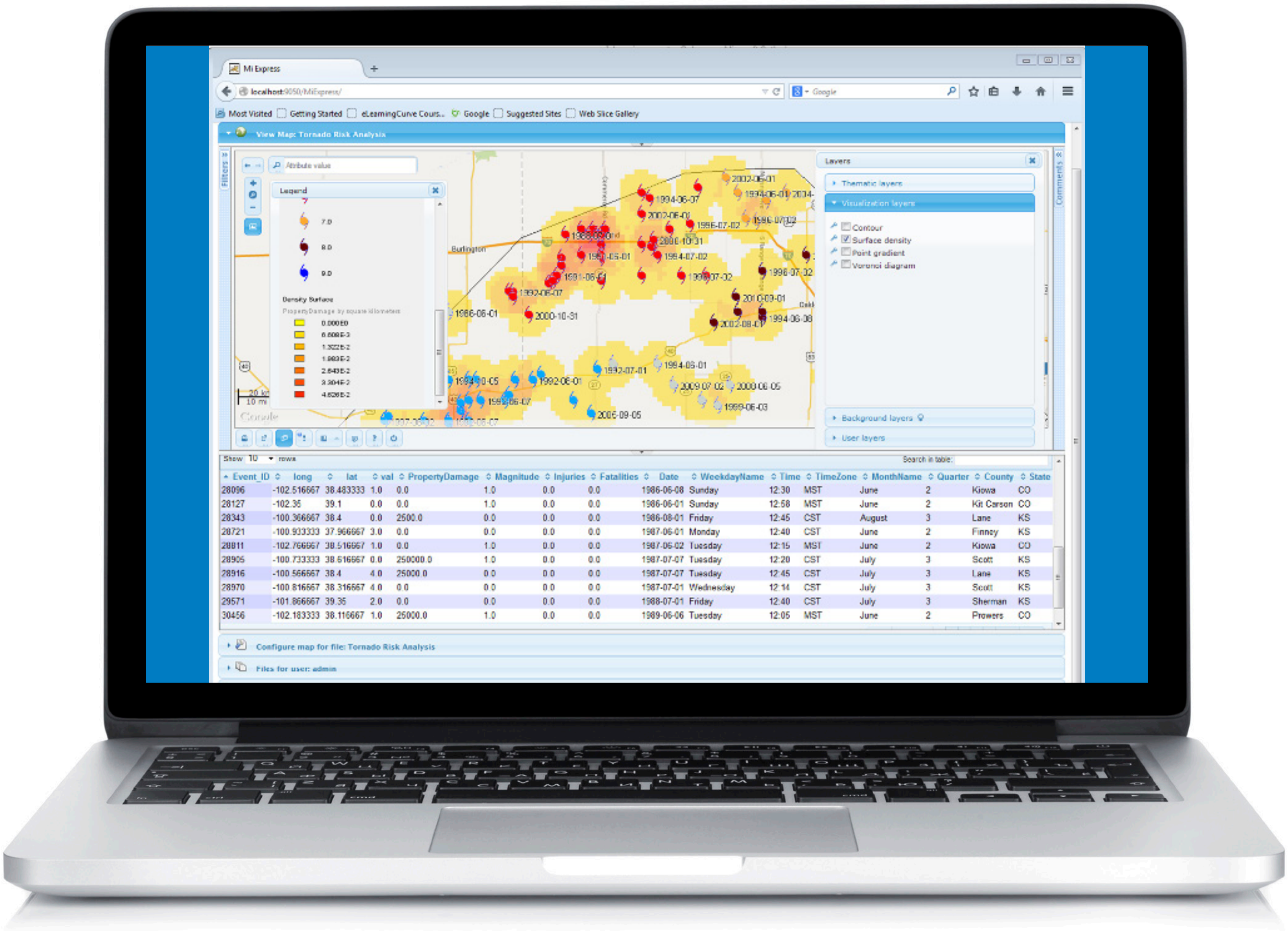
TIME DIMENSIONS ENABLE YOU TO ANALYZE YOUR DATA BY WORKDAYS VERSUS HOLIDAYS, WEEKDAYS VERSUS WEEKENDS, BY FISCAL PERIODS OR – IN THIS CASE – BY TORNADO SEASONS.

The web mapping application is used to automatically:

- Zoom to and display a map that includes the drive distance or radius boundary around the geocoded address ontop of Google Maps for context
- Generate individual value thematic map of tornado events by the R Density based spatial cluster
- Generate Surface Density Tornado Property Damage or Magnitude Heat Map
- Generate Bar Chart to show Tornado Spatial Clusters by Property Damage
- Display a table showing the attribute values for the mapped tornado events
- Provide interactive functionality to filter tornado events in the map, chart and table by tornado event attrbute, tornado map location, or bar chart spatial cluster selection



Tornado events within the study area are color coded by the R Density based spatial cluster and bar chart shows Tornado Spatial Clusters by Property Damage



Show 10 rows

Search in table:

Event_ID	long	lat	val	PropertyDamage	Magnitude	Injuries	Fatalities	Date	WeekdayName	Time	TimeZone	MonthName	Quarter	County	State
28096	-102.516667	38.483333	1.0	0.0	1.0	0.0	0.0	1986-06-08	Sunday	12:30	MST	June	2	Kiowa	CO
28127	-102.35	39.1	0.0	0.0	1.0	0.0	0.0	1986-06-01	Sunday	12:58	MST	June	2	Kit Carson	CO
28343	-100.366667	38.4	0.0	2500.0	0.0	0.0	0.0	1986-08-01	Friday	12:45	CST	August	3	Lane	KS
28721	-100.933333	37.966667	3.0	0.0	0.0	0.0	0.0	1987-06-01	Monday	12:40	CST	June	2	Finney	KS
28811	-102.766667	38.516667	1.0	0.0	1.0	0.0	0.0	1987-06-02	Tuesday	12:15	MST	June	2	Kiowa	CO
28905	-100.733333	38.616667	0.0	250000.0	1.0	0.0	0.0	1987-07-07	Tuesday	12:20	CST	July	3	Scott	KS
28916	-100.566667	38.4	4.0	25000.0	0.0	0.0	0.0	1987-07-07	Tuesday	12:45	CST	July	3	Lane	KS
28970	-100.816667	38.316667	4.0	0.0	0.0	0.0	0.0	1987-07-01	Wednesday	12:14	CST	July	3	Scott	KS
29571	-101.866667	39.35	2.0	0.0	0.0	0.0	0.0	1988-07-01	Friday	12:40	CST	July	3	Sherman	KS
30456	-102.183333	38.116667	1.0	25000.0	1.0	0.0	0.0	1989-06-06	Tuesday	12:05	MST	June	2	Prowers	CO

Configure map for file: Tornado Risk Analysis

Files for user: admin

A Surface Density Heat map of tornado property damage is shown below the tornado events

For more information call 800.327.8627 or visit us online: www.MapInfo.com

UNITED STATES/CANADA

800.327.8627/800.268.3282
pbsoftware.sales@pb.com
pbsoftware.canada.sales@pb.com

INDIA

+91.11.4219.5501
pbsoftware.india@pb.com

EUROPE/UNITED KINGDOM

+44.800.840.0001
pbsoftware.emea@pb.com

ASIA PACIFIC/AUSTRALIA

+61.2.9475.3500
pbsoftware.australia@pb.com
pbsoftware.singapore@pb.com

Pitney Bowes

RECYCLE PLEASE
recycleplease.org