Real-time Data Fusion, Georeferentiation and Analysis of Road Traffic Data reported by Multiple Sources

Ricardo Pinto[1], José Sousa[2], and Gabriel Pestana[1]

[1] INESC/IST, 1000-029 Lisboa, Portugal
[2]ESRI-Portugal,1600-131 Lisboa, Portugal

***Abstract*— Traffic congestions are nowadays a major problem in most countries. Businesses that use roads as means of transport are specially affected by this growing problem. Specifically for freight carries, traffic congestions are considered as a "serious" to "critically serious" problem for their business. Indeed, a timely delivery of merchandize is central to the quality of services provided, and consequently, to a long-term success. However, to guaranty a timely delivery of merchandize is a difficult task. Route optimization is fundamental, as are short delivery windows. But any significant alteration to traffic conditions can contribute to delays of one or more deliveries, proving chosen routes as wrong, and leading to non-compliance of pre-established delivery windows. Therefore, having up-to-date information about the current state of traffic conditions contributes to an improvement of carrier businesses by means of real time adjustments to transport routes and delivery plans. One approach is to provide drivers with up-to-date information about localized traffic congestions, combined with alternative route suggestions. For this purpose, this paper presents a methodology for a real-time data fusion, georeferentiation and analysis of road traffic data reported by multiple sources. A geographic and ontological data clustering process is introduced, followed by traffic impact estimation and data accuracy analysis algorithms. This methodology is being tested on a web-based prototype for fleet drivers in Portugal, which integrates updated traffic information with remote map and routing services.**

***Index Terms*— Accuracy Analysis, Alternative Route Suggestion, Data Clustering, Data Fusion, Georeferentiation, Impact Estimation, Real-time Traffic Management**

I. Introduction

The road freight transport has shown a regular increase over the last years. According to EUROSTAT, between 1996 and 2007 the quantity of transported goods has risen from 1.3 billion tkm (ton-kilometer) to 1.9 billion tkm, making the road transport the most used form of transportation, holding a 42% of the total of transported goods in Europe [1]. One of the main problems of this business is traffic congestions. According to [2], traffic congestions are considered as a "serious" to "critically serious" problem for freight carriers. They account for unreliable travel times, driver frustration and morale. Furthermore, traffic congestions are a leading cause for unnecessary fuel consumption and CO_2 emissions.

Freight carriers usually rely on route optimization algorithms and short delivery windows for the improvement of their business profits. But any significant alteration to traffic conditions can contribute to delays of one or more deliveries, proving chosen routes as wrong, and leading to non-compliance of pre-established delivery windows. Therefore, having up-to-date information about the current state of traffic conditions contributes to an improvement of carrier businesses by means of real time adjustments to transport routes and delivery plans. One approach is to provide drivers with up-to-date information about localized traffic congestions, combined with alternative route suggestions. For this purpose, this paper presents a methodology for a real-time data fusion, georeferentiation and analysis of road traffic data reported by multiple sources.

Traffic data may come from a wide range of different sources, ranging from common drivers that individually and proactively report incident and road congestion information, on one side of the spectrum, to road sensors and cameras specifically installed for the analysis of traffic status, on the opposite side. Considering the heterogeneity and multiplicity of possible data sources that might contribute with usable data, a normalization process is required in order to make sense of all the different data formats. Standards for describing road data are somewhat recent, being the Datex2 standard an up-to-date European reference for the interchange of road traffic data [3] and the UNETRANS a very thorough model for the organization of road related data, covering infrastructure, routes and dynamic data [4]. Unfortunately, a generalized adoption of standard protocols for the communication of traffic data has not happened yet. We decided to abstract ourselves from this drawback, assuming that all provided data follows a common format, in order to focus on the chosen content for this paper, without losing ourselves in data normalization related issues. Fig. 1 shows a proposal for the normalized data format, which resulted from an iterative simplification of a subset of the Datex2 specification.

The remaining text is divided into the following sections: first we introduce a process for the clustering and analysis of normalized highway traffic data provided by multiple sources with varying reliabilities; we subsequently explain how the resulting data can be geographically pinpointed using Teleatlas data and a process called dynamic segmentation; finally we present the interface of a developed web prototype, which integrates online map services with processed data, and provides best route calculation based on the current clustered traffic data.

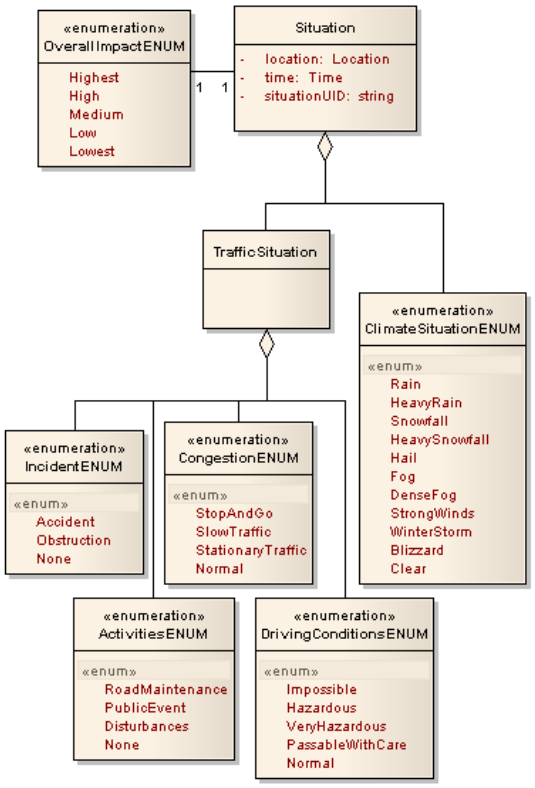

Figure 1 - Provided data model resulted from a simplification of a subset of the Datex2 specification. This subset is related to traffic events description.

II. Data Grouping

Considering multiple sources simultaneously feeding data to the system, leads us to reckon the possibility of more than one data record being referent to the same occurrence. In these conditions, fusing records not only promotes data unity but also contributes to a better ascertainment of the location of the registered events, as their respective individual veracity. Thus said, the problem we try to solve is how to discern which records should be grouped, and how to determine the real location of an occurrence composed by multiple records.

The first challenge we face is how to handle a large amount of data distributed across a vast geographic terrain. We base our solution on the divide-and-conquer paradigm. First off, records are separated into different sets. Each set refers to one highway and specific driving direction along it, which results in a pair of sets for each individual highway. Then, for each set containing traffic data, a grouping methodology is applied. This methodology has three main stages: (1) on the first stage, we separate the records by applying a variant of an agglomerative hierarchical clustering algorithm; (2) for each resulting group, an ontological analysis is made so as to determine which records are really related to the same occurrence; (3) finally, for each subgroup resulting from the previous analysis, two calculations are made: first the center of mass of traffic records is calculated - where mass represents the reliability of a record - so as to estimate the real location of the occurrence; second, the linear traffic extent of the event is estimated, for impact analysis and time span estimation. Each one of these stages will be subsequently described.

(1) The first stage consists of organizing the georeferenced records into separate groups based on their geographic proximity. An agglomerative hierarchical clustering algorithm is applied. When the algorithm starts, each record consists of a separate cluster. On each iteration, the two closest clusters are merged. The location of a cluster is calculated through the mean of the records' kilometric points the cluster holds. The algorithm stops when there is only one cluster left.

Formally, a set of clustered records $\boldsymbol{C}$, has a relative mean position $\boldsymbol{P_c}$ defined by:

$$\boldsymbol{P_c} = \frac{\sum_{i=0}^{n} p_i}{n}, \qquad p_i \in C$$

Equation 1

where p_i represents the kilometric point of the *ith* element contained in the set C.

Another way to represent the merged clusters is through a dendogram (view fig. 2). Each leaf node represents an individual record, whereas the remaining nodes represent clustered records, being the root node the only cluster left after the final iteration. A dendogram is a typical result of an agglomerative hierarchical clustering algorithm. After the dendogram is created, it remains only to choose how to cut the dendogram into subtrees, so that each subtree has only geographically close leafs. Being this somewhat subjective, it was determined that leafs considered as geographically close should have a standard kilometric point deviation below a certain specific threshold. Thus said, each node of the dendogram stores the standard deviation of all leafs of the subtree it is root:

$$\boldsymbol{S} = \sqrt{\frac{1}{n-1}\sum_{i=0}^{n}(p_i - \boldsymbol{P_c})^2}$$

Equation 2

Standard deviation then tends to increase from bottom nodes to the root node of the dendogram. With a depth first search starting from the root node, when reaching a node with a standard deviation below the defined threshold, this node is considered the root of a subtree with geographically close leafs, being the whole subtree cut from the dendogram (view fig. 3) and it's leafs stored in an individual group for subsequent analysis.

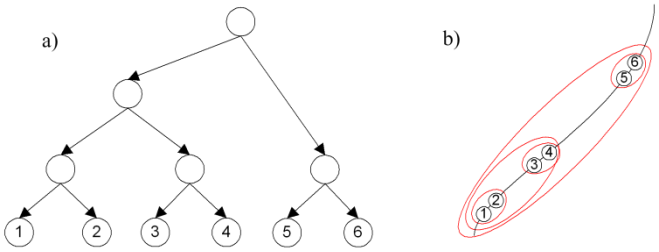

Figure 2 – Leaf nodes in (a) represent traffic occurrences. Nodes are grouped according to their geographic proximity. In (b), ellipses illustrate the grouped occurrences depicted in the dendogram.

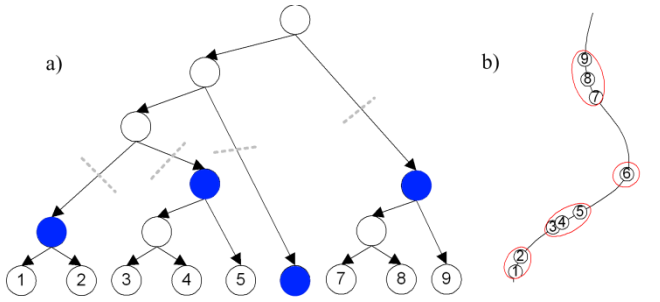

Figure 3 – A subtree is cut when its root node has a standard deviation below a specific threshold. Further analysis can thus be confined to each resulting subgroup of records. Root nodes are depicted in blue in (a). (b) illustrates the grouped records.

(2) The second stage consists of applying an ontological analysis to each one of the identified groups so as to determine which records are really related to the same occurrence. A depiction of the ontological model is shown in fig. 4. It is indicated that a road obstruction is the most generic description of an unintended traffic occurrence. Nevertheless, obstructions may have a more specific description, as modeled through the inheritance relationships. Moreover, they have a duration during which normal flow of incoming traffic may be disturbed, and also a location specifying the kilometric point and the side of the road where the accident occurred.

Slow traffic may always appear independently of traffic obstructions, most commonly due to congested roads on rush hours. As depicted in the diagram, "slow traffic" is a generalization of "queueing traffic", "stop and go" and "stationary traffic". It has also an estimated duration and location. The directionality of the Location property is the first characteristic used to distinguish geographically close traffic incidents. If they appear in opposite sides of the road, then most probably they are not related.

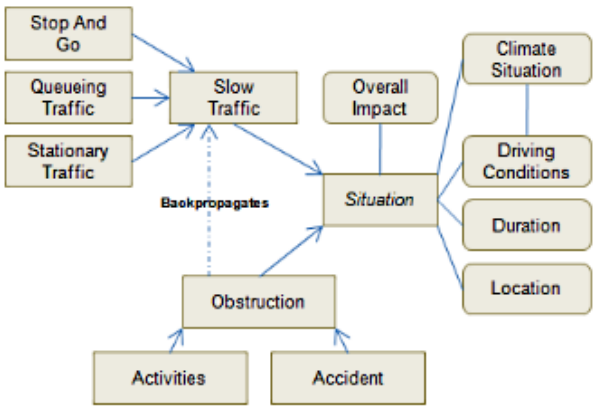

Figure 4 - Ontological model of the relationships between different incident descriptions. Inheritance and possible interchangeable relationships are depicted. Location and duration are mapped, as also a backpropagates relationship, which illustrates the effect that an accident may have on traffic further down along the road.

The generalization relationships are used to determine if two traffic records are in fact related to the same event despite of different descriptions. Note that an observed accident may be related to observed slow traffic some distance back down the road. This is modeled through the "backpropagates" relationship. Furthermore, previously reported accidents may be related to newly ones, depending on proximity and time span. Adequate proximity and time span thresholds are best when determined through data mining analysis of preexisting data; if not possible, simple intuitive assignments might suffice for satisfactory results.

The ontological classification algorithm has the following steps:

(a) First, divide the records into two separate groups based on the directionality of the events and store each individual record in a separate set.

(b) For both groups, determine which records represent generalizations of others. For those who are, determine if proximity and time span are below specified values. In case they are, merge both sets together.

(c) For each traffic flow record, verify, to a previously defined extent, if accident records exist further down the road. If records are found and all of them are on the same set, group them together. If not all of them belong to the same set, group the traffic flow record with the set which has the most recent records. For all traffic flow records that remained ungrouped to accident records, group them together.

(3) The third stage consists of determining, for each set resulting from the ontological analysis, the real location of the occurrence referred by each set's records, as also the estimated linear impact extent of each event along the road. There are three types of possible sets: sets with traffic congestion records only, sets with congestion and accident records, and sets solely with accident records. Each one of those is treated in a slightly different way.

For sets with congestion records only, the goal is to determine the real extent of the congestion, using the reported records as reference. Assuming that communicated records tend to have a higher frequency near the center of the traffic congestion, we presume that the distribution could be approximated to a normal distribution. Moreover, since the number of records should usually be rather small, the T-Student distribution presents itself as an even better approximation. Based on this assumption, we use the T-Student Distribution to calculate the confidence intervals of 90%, 80% and 70%. In this case, confidence intervals are used to answer the following question: assuming a specific set of records, calculate a kilometric interval that determines, with a confidence of X, where other congestion records related to this event might fall. Each confidence interval gives us the left and right side confidence limits, which in this case, represent kilometric points. Formally we have:

$$ConfLims_{1,2} = \boldsymbol{P_c} \pm A\,\frac{\boldsymbol{S}}{\sqrt{n}}$$

Equation 3

where A corresponds to the value from the T-Student table that is retrieved based on the provided confidence interval and number of degrees of freedom.

Pairs of limits translate to different extents of where the traffic congestion starts and ends, respectively (view fig. 5). These extents - stacked bottom up from 70% to 90% - give us a visual and kilometric perception of three levels of congestion: being the 90% confidence interval the one with the largest extent and representing a low level of congestion, as opposed to the 70% interval, holding the shortest extent and a high level of congestion.

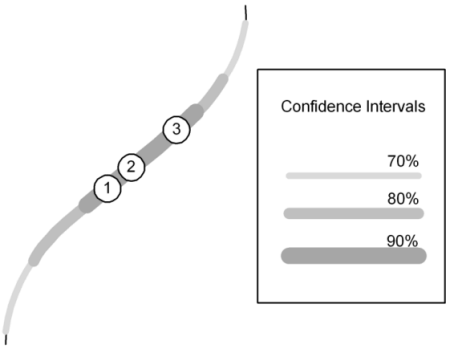

Figure 5 – Confidence Intervals are calculated for the kilometric points of three congestion reports. Depicted interval distances are only representative.

For sets with accident records only, the goal here is to estimate the actual location of the accident. Taking into account that each record has a specific reliability, we apply the center of mass theorem using the records reliabilities as their individual masses (view fig. 6). Formally:

$$\boldsymbol{P_r} = \frac{\sum_{i=0}^{n} p_i \, r_i}{\sum_{i=0}^{n} r_i}$$

Equation 4

where r_i represents the individual reliability of the ith record. The result is a kilometric point that best describes where the accident should be, based on the reliability weights of the reported records.

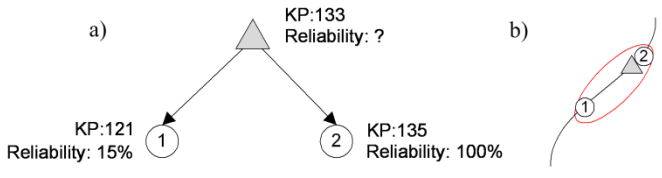

Figure 6 – The triangle in (a) represents the center of mass of the grouped records. The reliability of the new node is not yet defined. As illustrated in (b), the center of mass is closer to node 2 because of its higher reliability.

The last type of set is the one that contains congestion and accident records. The process applied here is similar to the ones above: first we divide the set into two subsets, on for congestion records and one for traffic records. For the congestion records we calculate the confidence intervals as described above, and for the accidents records the center of mass. Finally, the kilometric end limits of all confidence intervals are readjusted to end on top of the calculated center of mass of the accident records (view fig. 7).

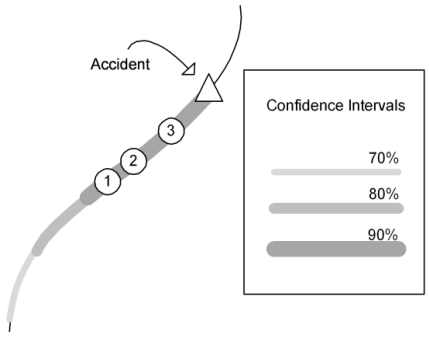

Figure 7 – End limits are readjusted to meet with the calculated center of mass of the accident records. Accident records are omitted for simplicity.

III. Dynamic Segmentation and Record Georeferentiation

Taking into account the normalized traffic data format provided in annex, data transmitted by external sources may come with three distinct description formats for their spatial locations: they may have only a textual description for their location; they may already be spatially referenced with an assigned geographic location; or they may have both. Each format needs to be treated in a specific manner. Highway traffic data with only a textual description for their location needs to be geographically pinpointed, a process called geocoding or georeferentiation. Oppositely, traffic data accompanied only with geographic coordinates needs the inverse treatment, formally named as reverse geocoding. This process consists of determining the textual description that best describes the location of a geographic coordinate, typically by determining the closest address to the given coordinate, or in our case, by determining its kilometric point along the highway. The third format needs only to be checked for concordance between geographic coordinates and textual description. In order to accomplish any of these tasks, geographically located geometric and route data of the highways in question needs to be readily available. Geographically located geometric data of road networks is available for purchase by various vendors. These data consist typically of short contiguous geometric features geographically positioned in respect to a specific coordinate system or Datum, being each one associated with a set of attributes related to driving rules and conditions, such as directionality, impedance, turning, velocity limits and so forth

[5][6]. Highways are described in a similar fashion too. However, these data only goes so far. In respect to determining a kilometric point of a specific location along a highway represented by geometric features, the available data, as it stands, does not suffice. The route that fully describes the highway must be determined, and its shape and location identified (one route for each driving direction along it). A process called dynamic segmentation solves this problem, and provides other advantages too.

Dynamic segmentation consists of a runtime computation of routes and route properties. Each route construction comprises an iterative process that runs through a set of previously identified calibrated linear features, assembling preselected properties along the way. In our case, distance measures were calculated (view fig. 8). The resulting route enabled us to calculate the metric distance of any chosen point along a preprocessed route. One of the main advantages of dynamic segmentation consists of the abstraction it provides from the features it is composed of. Changes to the underlying network may be made at any time, being those immediately reflected on the route's shape and properties. This hides the complexity of a route and encapsulates all route changes to one single place, making them innocuous to all route users.

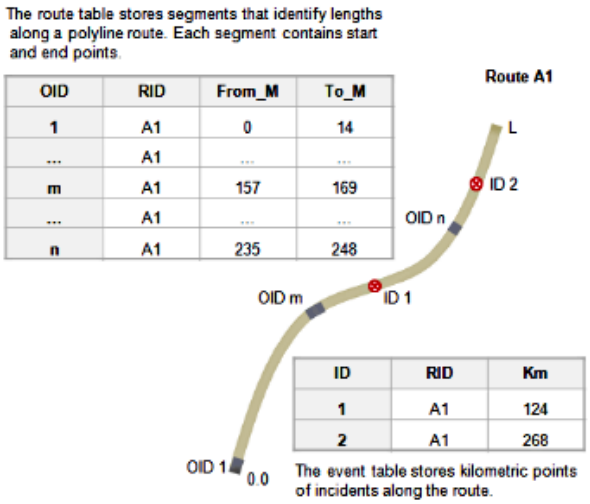

Figure 8 - The left table shows how route segments are stored in a dynamic segmentation table. The right table shows incident kilometric points ready for georeferentiation.

IV. Case Study

A hypothetical scenario was elaborated for one of the main Portuguese highways - named A1 -, which traverses half of the Portuguese territory, responsible for connecting the two main cities, Lisbon and Porto. The scenario was thought of as an especially hazardous day, with various accidents along the highway and various traffic and congestion reports. First off a layer of control data was created. It served as reference for the preparation of a set of simulated reports regarding traffic and congestion situations, and for subsequent comparison with the data produced by the system. The simulated reports were distributed on top of the reference data, being reliability typically correlated with the accurateness of the reports. All data, control, simulation and processed data were integrated in a geographic context and presented in a web interface (view fig. 9). This figure focuses on a specific highway segment with registered traffic and congestion data. The left image presents the control data layer, where red triangle represents where the accident occurred, and the yellow strip line the extent of congestion at the moment. The second image presents where the simulated traffic and congestion records were georeferenced. Triangles represent accident related records, whereas green dots represent congestion related records. The size of triangles is proportional to the reliability of the sources. The third image presents the processed data, with accident location and impact estimation.

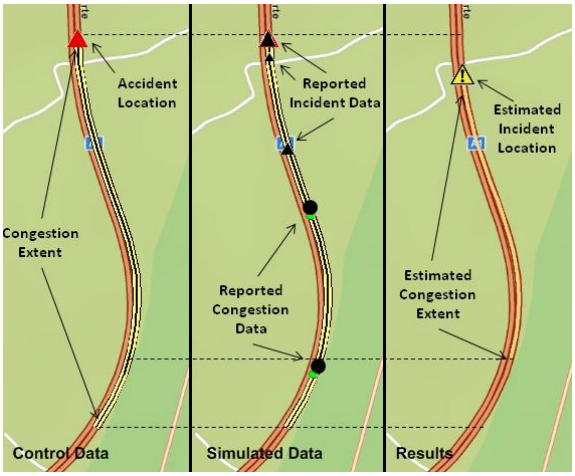

Figure 9 - Control data is shown on the left. The red triangle represents the accident location. The yellow line represents the extent of the congestion. Simulated data contains accident and congestion reports. Reliability of accident reports is depicted through black triangle size. On the right side, estimated accident location is shown, as well as congestion extent.

Routing capabilities were easily integrated in the system. For that purpose, ESRI European routing services were used. Each service call requires the specification of at least to stop points (start and finish), and an optional list of road barriers. The result of the service call consists of a route that traverses through all the specified points, avoiding the specified barriers, if located on the best route path. A routing example is shown in fig. 10.

The provided system service enables best route estimation, between two points, taking into account the collected traffic events. For each service request, first off, a list of all clustered traffic events is compiled. Then, for each clustered traffic event, its distance to the route start point is measured. The resulting distance is used as an estimate for the necessary travel time from start point to event. If estimated travel time is longer than the estimated end-time for the event, the event is removed from the list. The resulting list consists of all traffic events that are located close enough to be reachable from the route start point. This list is recalculated before every route request, and sent along as parameter to the remote routing service call.

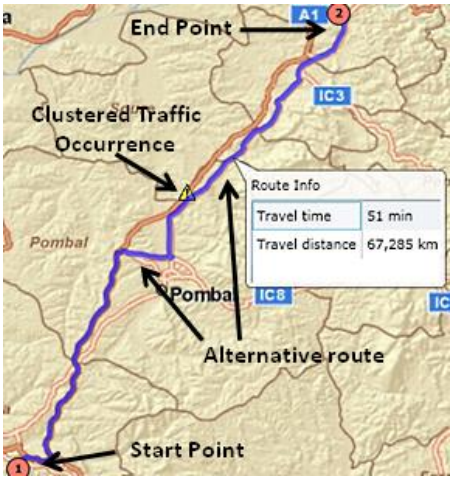

Figure 10 - Alternative route suggestion, presented on the right side of the image, is calculated after incident detection. Travel time and travel distance are both recalculated for the new route.

V. Conclusions

One of the main problems of freight carriers is related to traffic congestions, considered as a "serious" to "critically serious" problem for their business. One way to mitigate this problem is to collect current traffic data and use it as leverage for best route calculation. Contemplating current traffic data for route choice improves average driving speeds, which in turn contributes for the morale of the drivers, lower CO_2 emissions and more predictable delivery times. For this purpose, this paper presented a methodology for a real-time data fusion, georeferentiation and analysis of road traffic data reported by multiple sources. A method for the fusion and georeferentiation of real-time highway traffic data, so as to provide a geographically enabled, time and impact aware view of current traffic status in highways was described. A grouping algorithm based on hierarchical clustering and ontological analysis was presented. Center of mass theorem and confidence intervals were used for the estimation of traffic event locations and respective cue lengths. We used a dynamic segmentation algorithm for route definition and subsequent geocoding and reverse geocoding operations. All data was integrated through SQL Server 2005 and ESRI ArcGIS Desktop 9.3.1. The resulting data was published through ArcGIS Server 9.3.1, and consumed by a Silverlight web application. Feature identification and routing tasks were easily integrated. Routing took into account clustered traffic data for best route estimation.

Acknowledgment

R. Pinto thanks ESRI-Portugal for their active support on the elaboration of this paper and also would like to thank his Master's Supervisor, G. Pestana, for his very insightful reviews and comments that helped shape this document.

References

[1] EUROSTAT Statistical Books: Panorama of Transport, pp. 56-57. 2009 Edition.
[2] Impacts of highway congestion on freight operations: perceptions of trucking industry managers. Institute of Transportation Studies, University of California. 2001
[3] European Commission - Directorate General for Transport and Energy: Datex II (version 1.0). User Guide (2006)
[4] Curtin, K., Noronha, V., Goodchild, M., Grisé, S.: ArcGIS Transportation Data Model (UNETRANS). Technical Report (2003)
[5] Gan, A., Liu, K.: A Robust Dynamic Segmentation Tool for Highway Safety Analysis. In: Applications of Advanced Technology in Transportation, pp.263.268. Proceedings of 9th International Conference, Chicago (2006)
[6] ESRI White Paper: Linear Referencing and Dynamic Segmentation in ArcGIS 8.1. Technical Report (2001)