Development and evaluation of an intelligent fleet management system for city logistics

Vasileios Zeimpekis, George M. Giaglis
Department of Management Science & Technology
Athens University of Economics & Business
Athens, Greece
{vzeimp, giaglis}@aueb.gr

Ioannis Minis
Department of Financial Management & Engineering
University of the Aegean
Chios, Greece
i.minis@fme.aegean.gr

Abstract
The routing plan of a delivery fleet is usually developed a priori and provides an optimal (or near optimal) way of serving customers by taking into account certain constraints, such as delivery time windows. However, such plans may not cope adequately with the dynamics of a city logistics environment, in which unexpected events (e.g. traffic congestions) often occur during delivery execution. We present the development and evaluation of a real-time fleet management system that handles such unforeseen events. The system monitors the delivery vehicles in real time, detects deviations from the initial distribution plan, and adjusts the schedule accordingly by suggesting effective rerouting strategies. The system has been tested in simulation environment and in real-life cases and the results show that delivery performance is enhanced significantly and customer satisfaction is improved.

1. Introduction
Freight distribution arguably accounts for a significant portion of the total costs of logistics [1]. Efforts to minimize distribution costs typically focus on the development of near-optimal distribution plans using various types of vehicle routing algorithms. Urban distribution, however, is more susceptible to unexpected costs and delays that arise during the execution of a delivery plan due to unforeseen adverse conditions. One can distinguish three main types of such incidents:

1. Incidents originating from the clients served: Typical examples include order cancellation, delivery time changes, new customer requests and lack of unloading or parking space at the customer site.

2. Incidents related to the road infrastructure and environment: Traffic congestion, road construction, flea markets, rain or snow and so fourth.

3. Incidents related to the delivery vehicle: Typical examples include accidents and/or mechanical failures.

The use of an initial distribution plan, although necessary, is by no means sufficient to address such unexpected events that may have adverse effects on system performance. Recent advances in mobile and positioning technologies have allowed the development of fleet management systems that enable distribution organizations to monitor their fleet in real-time and to improve the performance of the delivery network by mitigating some of the aforementioned problems. However, the systems based on these technologies are not typically designed to address unforeseen events in a systemic fashion.

Recent work focused on incident handling systems that have addressed the case in which a new customer request appears during the execution and must be fulfilled in a specific time period [10, 11]. This paper addresses another practical aspect of significance which relates to the dependency of travel and service times on the time of the day [8]. Most available incident handling fleet management systems [9, 12, 13] assume that the travel and service times are constant throughout the day or use simple procedures to adjust them, such as multiplying factors associated with different periods of the day. However, the performance of these approaches depends on the validity of the assumptions introduced, resulting, in many cases, in suboptimal or even infeasible situations [8].
The remainder of the paper is organized as follows. Section 2 presents the main requirements for dynamic incident handling in an urban distribution environment. Section 3 provides an overview of the architecture of the real-time fleet management system, whereas Section 4 analyses the operation of the system and shows results from its implementation in a third party logistics (3PL) company. Concluding remarks as well as the main benefits from the use of the system are included in Section 5.

2. User Requirements

In order to define the main attributes of the real-time fleet management system, a three-phase user requirements elicitation process [5] has been followed. The first phase (Phase 1) comprises an analysis of the transportation and logistics industry taking into account insights from telematics industry specialists. Phase 2 comprised the core (quantitative) analytical phase and included a survey (on-line questionnaires) focusing on the needs for real-time fleet management services with 73 responses. The output of this phase led to the selection of desired services. Finally, in Phase 3 we verified the results of Phase 2, via qualitative in-depth interviews with 15 logistics directors from Greece.

The results from this process revealed that existing fleet management systems can fulfill a subset of the requirements expressed by logistics and distribution organizations. The requirements included in this subset are:

- Monitoring of the geographic location of vehicles in real time: This is a typical requirement, which is fulfilled through the use of satellite and terrestrial communication systems.
- Generating a posteriori reports of vehicle and distribution system performance.
- Generating proof-of-delivery (POD) statements upon completion of load delivery.

However, a number of other requirements cannot be effectively met by existing systems. These include:

- The ability to intelligently reroute a truck that is delayed and will miss delivery time windows, thus minimizing the adverse impact on overall delivery performance.
- The ability to deal with breakdowns that immobilize a vehicle, by rerouting another nearby vehicle (or vehicles) that has both adequate load capacity and time availability to reach the immobilized vehicle, load its goods, and continue the delivery.

The aforementioned requirements were synthesized into design specifications for the real-time fleet management problem.

3. Overview of the real-time fleet management system

3.1 System Architecture

The proposed system (Figure 1) comprises three subsystems, back-end, wireless communications, and front-end, which are discussed below.

![Architecture of Real-Time Fleet Management System for Dynamic Incident Handling](image)

The back-end system incorporates typical components of a fleet management system such as a) a Geographical Information Module (GIM), responsible for managing cartographic information, b) a Data Management Module (DMM) that incorporates data related to clients, vehicles, and distribution schedules, c) a Control Centre User Interface, and d) an innovative Decision Support Module (DSM); the latter is responsible for dynamic incident handling.

The wireless communication sub-system consists of two parts: a) The mobile access terrestrial network, which is responsible for the wireless interconnection of the back-end system with the front-end on-board devices, and b) the positioning system, which is responsible for vehicle tracking. We have used a) GPRS network for terrestrial data transmission, since it can support efficiently real-time transfer of data, and b) GPS for vehicle...
tracking, a globally available, free-of-charge system.

The front-end system consists of the telematic equipment that supports real-time communication and data processing (Vehicle On-Board System), as well as a portable data terminal for use by the driver (Vehicle User Interface).

### 3.2 Decision Support Module (DSM)

DSM computes and proposes all required real-time adjustments to the delivery plan, in order to meet preset goals taking into account the system’s dynamic state. It implements the incident handling mechanism that consists of three stages: a) Monitor & Detection b) Delivery Trip Projection and c) Decision Making & Rerouting. The Monitor & Detection component is responsible for monitoring the dynamic state of each vehicle in the fleet (geographical position, speed, and inventory) and detecting possible deviations from the original delivery plan. The Delivery Trip Projection component uses a travel time estimation method in order to identify whether the remaining time from current position to the clients not yet served is less than or equal to the upper limit of the related time windows. The Decision Making and Rerouting component is activated when a time delay is detected and vehicle rerouting is required based on the Trip Projection calculations. This module solves, in real-time, appropriate mathematical programming models that describe the two cases mentioned above: a) rerouting of a delayed vehicle, and b) rerouting of one or more vehicles in the case of vehicle immobilization due to a breakdown. The first case is overviewed in this paper.

The mathematical programming formulation developed to model the delayed vehicle rerouting problem has strong similarities with the so-called orienteering problem (OP), a generalization of the extensively studied Travelling Salesman Problem (TSP). The OP concerns a sportsman that visits a set of geographical sites collecting a certain prize from each site. The objective is to maximize the total collected prize; the task is constrained by a time (or distance) upper limit. The first to formulate the OP was Tsiligirides [14], who proposed two heuristic solution methods, as well as some, widely used, benchmark problem instances. Chao et al. [15] proved that OP is NP-hard, and proposed an efficient three-step iterative heuristic.

In addition to the time upper limit, there may be time window constraints in one or more clients; that is the start of service in a client cannot begin before the client’s time window opens, or after the client’s time window closes. Note that there are also significant efficiency requirements that should be satisfied in any real-time implementation. We have developed an efficient algorithm that has been also compared with known solutions of standard problems (from the literature) and was found that the latter provides a near-optimal solution to this problem. The basic idea is that the route is constructed iteratively by inserting clients based on a roulette wheel function that uses a client ‘desirability’ metric for the client insertion. The desirability metric strikes a balance between a) a parameter defined as the profit “pi” that the vehicle collects by visiting client i in relation to b) the distance spent to visit client i. The steps of the algorithm are as follows:

- **STEP 1** – From current point (i.e. the current point of the procedure is the location of the vehicle at the time rerouting commences) check the insertion feasibility of all unvisited clients. Insertion feasibility includes the satisfaction of time constraints (both for clients’ time windows and the depot return deadline).
- **STEP 2** - For all feasible for insertion clients, define a metric that quantifies the ‘desirability’ to insert those clients in the route and construct a vector comprising of the four most desirable clients. Finally, choose stochastically based on a roulette wheel function- the client to be inserted in the route.
- **STEP 3** – Repeat steps 1 and 2 until no more clients can be feasibly inserted in the route or until the client set is empty. At this point one feasible solution has been obtained.
- **STEP 4** – Repeat steps 1, 2 and 3 for M iterations in order to get M feasible solutions
- **STEP 5** - Select the best route out of the M feasible solutions. The best route is the one that gives the maximum cumulative income to the vehicle

The algorithm is strongly influenced by the S-algorithm of Tsiligrides [14]. However, the introduced time window constraints allow for further investigation of the ‘desirability’ metrics of the clients as well as the stochastic choice of the next client to be inserted in the network. These two parameters play an important role in the quality of the solution. The algorithm has been implemented in Matlab 7.0, compiled to C and inserted into the
3.3 System operation

The algorithms for dynamic incident handling were embedded in the fleet management system. This mechanism was controlled via a user interface in the control centre that is shown in Figures 2a, and b. We have also developed a special software application in the on-board terminals that allowed the driver to have a bi-directional communication with the control centre so as to receive rerouting decisions, provide proof-of-delivery information, as well as alerts in case of unforeseen events.

A typical operation of the real-time fleet management systems is as follows. Figure 2a depicts the control centre user interface at an initial stage, when the travel estimation technique does not detect any deviation from the initial delivery plan (the column, which depicts the estimated arrival time for each customer, is green). After a vehicle has served a number of customers (Figure 2b), the system detects several time violations for non-served customers (certain cells of the corresponding column are red), and proposes a rerouting plan (i.e. a different way for visiting the remaining customers).

The new delivery plan is transmitted to the driver through the on-board terminal (Figure 3).

4. System evaluation

The proposed real-time fleet management system, has been tested in two Greek companies: a large third-party (3PL) delivery company (DIAKINISIS S.A.) and a food product manufacturer (P.G. NIKAS S.A.) that operates its own truck fleet. The case of the 3PL company is described below.

4.1 Description of company’s urban distribution operations

Diakinisis S.A. is one of the largest third party logistics (3PL) companies in Greece. Its core business focuses on the storage, order management, invoicing, and distribution of goods for a large number of commercial and manufacturing companies. It is situated about 15km from the Athens and Piraeus city centres where more than 60% of its customers are concentrated. Every day, more than 150,000 Kg of goods are distributed to an average of 1500 clients, located at distances ranging from 3 to 40 km (in the Attica Prefecture) from the company’s main warehouse. The orders are delivered using an outsourced fleet of 80 vehicles. The company uses a customised software solution for the routing of these vehicles.

Due to the highly congested urban environment of Athens and Piraeus, the company faces various problems from unexpected events (mainly travel and service time delays), that may adversely affect the delivery process (e.g. violations of delivery time windows). Currently, when a delay occurs, interventions are often performed through voice communication between the driver and the dispatcher. Oftentimes the effectiveness of these interventions is limited, since there is no systemic way of taking into consideration the multitude of parameters involved, such as the importance of the
remaining clients, time window restrictions, and so on. Thus, the Diakinisis case is an ideal test venue for the proposed real-time fleet management system.

4.2 Design of pilot testing

The design of the pilot test was based on the design parameters presented in Table 1. According to previous studies [2, 3] these parameters affect the performance of the urban freight delivery process.

<table>
<thead>
<tr>
<th>Table 1. Design parameters of the field experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Design parameters</td>
</tr>
<tr>
<td>-------------------</td>
</tr>
<tr>
<td>Area</td>
</tr>
<tr>
<td>Traffic</td>
</tr>
<tr>
<td>Type of time windows</td>
</tr>
<tr>
<td>Number of time windows</td>
</tr>
<tr>
<td>Range of time windows</td>
</tr>
</tbody>
</table>

We have adopted a field experiment approach to perform the pilot testing [4]. To examine the robustness of the system with respect to a number of factors, a $2^k-m$ fractional factorial design was executed with respect to the aforementioned design parameters. Factorial experiments (including fractional factorial ones) are systematic ways to perform an experimental investigation that yields all statistically significant effects of all factors and their interactions [7]. It is noted, however, that a case with 5 factors and 2 levels (as in our case) requires $2^5 = 32$ runs times the number of replicates per run for a complete factorial experiment. In order to reduce this large number to a practical one, the one-quarter fractional factorial design was used corresponding the $2^{5-2}$ (i.e. $2^3$) treatment combinations. The eight test cases are presented in Table 2.

<table>
<thead>
<tr>
<th>Table 2. Test case scenarios</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test case</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
</tbody>
</table>

Area: Sub-urban (S), Urban (U)
Traffic: Heavy (H), Light (L)
Type of time windows: Driver’s shift (D), Customer’s restriction (C)
Number of time windows: Small (S), High (H)
Range of time windows: Relaxed(R), Tight (T)

Having developed the experimental design, two types of system evaluation were conducted: a) extensive simulation tests with 30 replicates (i.e. repetitions of exactly the same experiments, which are used to access the variability intrinsic to the results and thus, deduce statistical significance) per combination for a total of 240 runs and b) actual tests with one replicate per combination using two trucks.

4.2.1 Simulation Testing

During the simulation tests we have simulated the entire system with all its functionalities. That means: a) the monitoring process, b) the trip projection that incorporates the travel prediction methods, c) the decision support module that incorporates the rerouting algorithms as well as d) the user interfaces (both in the control centre and at the vehicle). For each scenario we have engaged 2 vehicles. All the above were simulated in an integrated manner so as to be as similar as possible to real-life testing of the system.

As mentioned in Section 3.2, in case of a delayed vehicle, the new plan may exclude one or more clients from the route (i.e. a set of customers), in case it is not feasible to serve all clients due to the accumulated delay. In order, however to prioritize the customers to serve such a case, a weight (i.e. customer importance) was given to each client. The latter was quantified by a factor (from 1-less important to 10-very important) according to the type of the customer and its importance. Based on this, in order to evaluate the customer service achieved by each vehicle of a certain test, we calculated the ratio derived by the sum of weights of customers visited by each truck over the total weight of all customers in the specific delivery plan.

\[
CS_j = \frac{\sum_{i \in S_j} w_i}{\sum_{i \in T} w_i} \times 100 \quad j=1,2,\ldots,n
\]  

where

$S_j$ is the set of customers visited by truck $j$
$S_T$ is the entire set of all customers in the route
The route that achieves the highest $CS_j$ is the preferred one. In many cases, this route may contain a lower number of high importance customers. However, a route with a higher $CS_j$ score and a higher number of customers served is clearly superior. Figure 4 shows the customer service (quantified by $CS_j$) achieved for each test case.

As it can be seen, in all test cases, when the directions provided by the real-time fleet management system were followed, a higher performance was achieved. Especially in cases where simulation tests were conducted in urban areas or included a lot of customers with time delivery restrictions, the proposed system had a superior performance. Thus, it was concluded that the use of real-time fleet management system affects the performance of urban freight delivery. The actual tests to confirm these results are described below.

4.2.2. Live deployment

In the live deployment, two vehicles were engaged simultaneously in each test. The first vehicle attempted to perform the daily schedule as originally planned, whereas the second vehicle followed the directions provided by the real-time fleet management system. Both vehicles were originally planned to serve precisely the same delivery points under identical circumstances following identical routes. In order to force delays, the delivery period was artificially set to be less than the time usually required for a vehicle to complete the delivery plan. For instance, the delivery period was set between 8.30 am-12.30 pm, although the time usually needed to complete the delivery plan ranges from 8.30 am to 2.00 pm. In this way, we forced the incident handling system to reroute the second vehicle.

4.3 Results from pilot testing

Figure 5 shows the customer service (quantified by $CS_j$) achieved for each test case. For all test cases, the vehicle that followed the new route designed by the real-time fleet management system (Vehicle B), provided higher customer service. As it can be seen the results are close to the simulation results presented in figure 4.

Table 3 (p. 9) summarizes the results for each test case in detail. For each test case this table provides the initial number of customers, the number of visited customer by each truck, the importance of visited customers, the total customer services and the performance achieved in each test case. The performance difference (Table 3) in $CS$ between the new plan and the initial one is 25 points, that is an improvement of a factor of 1.39 over the initial plan.

An important finding from this pilot test has been the impact of time windows on the system’s performance. Indeed, in the first three cases (Tests
1, 2 and 3) in which only the restriction of the driver’s shift has been applied, the total number of served customers by Trucks A & B was almost equal. In Tests 4, 5, 6, 7 and 8, which included customer time windows (in addition to the driver’s shift), the system performed better (in terms of customers served) and succeeded in reducing time window violations. The number of customers in the initial plan is another important issue. The higher the number of customers, the better the performance of the system becomes. This occurs because when a large number of initial customers that have certain delivery constraints exists, and an unexpected event occurs, it is difficult for a driver that has no guidance from the proposed system to choose, which customers to visit and at which order. Results from cases 1, 5, and 7 (all of which include a large number of initial customers), show that Truck B visited most of the customers and, in particular those with higher importance.

5. Conclusions

This paper proposed a system for dynamic incident handling with emphasis on vehicle delays. The proposed system increased customer service by 25% and improved the execution of urban freight deliveries. This was concluded both from simulation testing and live deployment of the system. Indeed the 3PL company, we cooperated with, recognized various strategic benefits such as real-time decision making, increased customer service, as well as operational benefits such as reduction of fleet management cost and dynamic incident handling capability.

The ideas presented here may be extended to emergency services, couriers, rescue and repair services as well as taxi cab services. In each case, the main characteristics of the system as well as its functionality will be targeted to each application domain. For example, in the case of courier services, other requirements could be taken into account, such as the actual time for pickup and delivery. Of course, the generic properties of the real-time fleet management system (i.e. incident handling by using dynamic travel prediction methods and rerouting algorithms), as specified in this research, will still be applicable.

Acknowledgments

The authors gratefully acknowledge the Hellenic Ministry of Education (Herakleitos PhD Fellowship) and the Hellenic General Secretariat of Research & Development for partially funding this research under the project entitled “Mobile Real-Time Supply Chain Execution”.

Figure 5. Customer Service for all cases in DIAKINISIS S.A.
References


Table 3. Results from the pilot testing in DIAKINISIS S.A
(Truck A follows the predefined delivery schedule whereas Truck B follows the directions provided by the real-time fleet management system)

<table>
<thead>
<tr>
<th>Test case</th>
<th>Type of windows</th>
<th>Number of scheduled customers</th>
<th>Number of visited customers</th>
<th>Importance (weight) of visited customers</th>
<th>Total distance traveled (km)</th>
<th>Total Customer Service (%)</th>
<th>Performance Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Driver's Shift</td>
<td>Truck A &amp; B</td>
<td>25</td>
<td>14</td>
<td>53,32</td>
<td>59</td>
<td>22</td>
</tr>
<tr>
<td>2</td>
<td>Driver's Shift</td>
<td>Truck A &amp; B</td>
<td>9</td>
<td>5</td>
<td>36,32</td>
<td>56</td>
<td>28</td>
</tr>
<tr>
<td>3</td>
<td>Driver's Shift</td>
<td>Truck A &amp; B</td>
<td>9</td>
<td>5</td>
<td>32,14</td>
<td>60</td>
<td>45</td>
</tr>
<tr>
<td>4</td>
<td>Customer's restrictions</td>
<td>Truck A &amp; B</td>
<td>7</td>
<td>4</td>
<td>31,56</td>
<td>60</td>
<td>25</td>
</tr>
<tr>
<td>5</td>
<td>Customer's restrictions</td>
<td>Truck A &amp; B</td>
<td>25</td>
<td>15</td>
<td>59,80</td>
<td>65</td>
<td>23</td>
</tr>
<tr>
<td>6</td>
<td>Customer's restrictions</td>
<td>Truck A &amp; B</td>
<td>6</td>
<td>5</td>
<td>39,30</td>
<td>74</td>
<td>28</td>
</tr>
<tr>
<td>7</td>
<td>Customer's restrictions</td>
<td>Truck A &amp; B</td>
<td>21</td>
<td>13</td>
<td>46,67</td>
<td>67</td>
<td>15</td>
</tr>
<tr>
<td>8</td>
<td>Customer's restrictions</td>
<td>Truck A &amp; B</td>
<td>5</td>
<td>3</td>
<td>27,46</td>
<td>70</td>
<td>15</td>
</tr>
<tr>
<td>TOTAL</td>
<td></td>
<td></td>
<td>107</td>
<td>64</td>
<td>326,57</td>
<td>64</td>
<td>25</td>
</tr>
<tr>
<td>AVERAGE</td>
<td></td>
<td></td>
<td></td>
<td>387</td>
<td>357,56</td>
<td>80</td>
<td></td>
</tr>
</tbody>
</table>