Lecture Notes in Logistics

Dynamics in Logistics

Third International Conference, LDIC 2012 Bremen, Germany, February/March 2012 Proceedings

Bearbeitet von
Hans-Jörg Kreowski, Bernd Scholz-Reiter, Klaus-Dieter Thoben

1. Auflage 2013. Buch. xiv, 580 S. Hardcover
ISBN 978 3 642 35965 1
Format (B x L): 15,5 x 23,5 cm
Gewicht: 1051 g

Fault Detection in Dynamic Vehicle Routing Operations

Antonio G. N. Novaes, Edson T. Bez and Paulo J. Burin

Introduction

The explosive growth in computer, communication, and information technology in recent years, together with dramatic changes in organizations and markets, have opened new forms of operating manufacturing and transport activities in an integrated and collaborative way (Goel 2008). To optimize performance, supply-chain functions must operate in a coordinated manner. But the actual circumstances observed in these operations make it difficult to implement it in many instances. Truck breakdowns, road traffic congestions, labour absences, customer's cancel or postponement of orders, among other random events, generate deviations from the basic plans. Thus, the management of these integrated systems must be performed in a dynamic way, revising the plans and schedules whenever it becomes necessary and when system failures require corrective interventions.

An unbalanced and unstable integration of manufacturing and transport systems can impair the competiveness of supply chains. This integration is even more relevant along global supply chains due to longer transport lead-times and the network complexity of manufacturing processes. Nowadays, production and transport scheduling are still carried out sequentially, due in part to their

A. G. N. Novaes (✉)
Federal University of Santa Catarina, Rua Beija-Flor 112,
Florianopolis 88062-253 SC, Brazil
e-mail: novaes@deps.ufsc.br

E. T. Bez
Univali – Itajaí Valley University, São José, SC, Brazil
e-mail: edsonbez@gmail.com

P. J. Burin
Federal University of Santa Catarina, Florianópolis, SC, Brazil
e-mail: pburin@gmail.com

H.-J. Kreowski et al. (eds.), *Dynamics in Logistics*, Lecture Notes in Logistics,
DOI: 10.1007/978-3-642-35966-8_2, © Springer-Verlag Berlin Heidelberg 2013

complexity and current lack of appropriate heuristics for supporting a desirable integration at the operational level. Especially within dynamic environments, production and transport systems must be properly integrated so that efficiency, responsiveness and flexibility could be achieved and sustained. Specially, the vision of a supply chain as synchronized systems of material and information flows requires that transport capabilities, level of utilization of resources and transit lead-times be appropriately employed in order to get the most effective and sustainable production scheduling (Frazzon et al. 2010).

The increasing complexity of technological processes, the availability of advanced sensor devices, and the existence of sophisticated information processing systems, have opened the way to detect abrupt as well as latent changes in some characteristic properties of a system. Fault detection and diagnosis applied to automatic control of technical systems have been extensively investigated in the literature (Basseville and Nikiforov 1993; Isermann 1997, 2005; Simani et al. 2010), but its concepts and methods have not been extended so far to services such as freight transport and logistics. The objective of fault detection in integrated manufacturing and logistics systems is to anticipate counteractions in order to avoid malfunctions and unexpected interruptions. Many fault monitoring problems can be seen as the problem of detecting a change in the parameters of a dynamic stochastic system. Model-based fault diagnosis methods are designed to detect abnormal situations confronting real data with modelling estimates. It is assumed that a discrepancy signal is directly or indirectly linked to a fault. Care must be taken to bypass model mismatches or noise in real measurements, which can erroneously be seen as a fault, giving rise to a false alarm in detection. These considerations have led to research efforts toward robust methods, with the aim of minimizing such drawbacks (Simani et al. 2010).

Dynamic vehicle routing problems (*DVRP*) have received increasing attention among researchers (Psaraftis 1995; Larsen 2001; Ribeiro and Lorena 2005; Larsen et al. 2008; Golden et al. 2008; Novaes et al. 2011). These problems are usually related to efficiently assigning vehicles to tasks, such as picking-up components from *OEM* facilities in a row, delivering cargo, or accomplishing other services in a previously defined order, so that tasks are completed within a certain time limit and vehicle capacities are not exceeded (Figliozzi 2007, 2010). But in large and congested urban areas, particularly in developing countries, transport operators tend to assign larger numbers of visits to their vehicles in order to increase revenue. This often leads to non-performed orders at the end of the daily cycle-time, impairing the logistics service level and postponing tasks to next day, or even later. This happens because, due to the volatile traffic conditions and the great number of random variables along the route, the vehicle cycle-time usually shows great variability. But even assuming that the fleet of vehicles has been well dimensioned, there are situations in which the traffic becomes exceptionally over-congested due to severe accidents, unpredictable public transport strikes, abnormal weather conditions, etc. However, when operating in a production schedule, comprising the pick-up of components from several *OEM* facilities in a row and carrying them to an assembler company, the vehicle has to accomplish its tasks within a pre-established JIT service level.

Traffic information systems, which have been installed in some large cities of the world, tend to increase the flow of vehicles by allowing higher vehicular speeds and by offering less-congested alternative routes to drivers (Fleischmann et al. 2004). The benefits of using such traffic navigational systems in connection with vehicle routing in congested urban areas cannot be denied. But in developing countries, the required large investments to install such systems often forbid its extensive use. One of the objectives of this work is to show that simple dynamic vehicle routing procedures can dramatically improve the logistics performance of the servicing system. With an on-board computer, a fault-detection software, and simple telematics devices linking the vehicle to nearby collaborative agents (other vehicles and the central depot), it is possible to attain better performance levels. By analysing vehicle operational data at specific regeneration points along the route, it is possible to anticipate the occurrence of unperformed tasks, emitting information to other agents (vehicles, central depot), and transferring part of the tasks to other participants. With this procedure the occurrence of unperformed tasks at the end of a vehicle cycle-time can be dramatically reduced.

In a *DVRP*, not all information relevant to the planning of the routes is known by the planner when the routing process begins, and information may change after the initial routes have been constructed (Larsen et al. 2008). In the application object of this work, although the vehicle service is fully planned in advance, the possible transfer of tasks to other agents and the eventual reprogramming of visits lead to changes in the routing process, thus characterizing a dynamic behaviour (Psaraftis 1995; Larsen 2001).

The Static Routing Problem

When modelling dynamic logistics problems it is necessary to quantify a number of random parameters to be used in the main mathematical model. Larsen (2001) dedicates a full chapter of his work to the computer simulation of such data. The reason is that real-life datasets are very often not available in such detail and accuracy as to support a thorough investigation of dynamic problems. Then, randomly generated data and simulation are frequently used when designing dynamic logistics systems. We make use of such a technique to analyse important parameters related to the routing vehicle cycle.

Throughout the paper an empirical case study will be examined based on a real-life Brazilian urban scenario. Let us consider an operating district of area A containing n *OEM* suppliers. The vehicle assigned to the district leaves the depot early in the morning, goes to the assigned district, performs the collecting service visiting the *OEM* suppliers, and goes to the assembler plant when all tasks are completed, or when the maximum allowed working time per day is to be reached, whichever occurs first. This complete time sequence makes up the vehicle cycle. In some practical circumstances more than one tour per day can be assigned to the same truck. This implies extra line-haul costs, but depending on the cargo

characteristics, vehicle size restrictions and other factors, multiple daily tours per vehicle might sometimes be appropriate. For the sake of simplicity, we assume that the vehicles perform just one cycle per day. The model can be easily modified to take into account multiple daily cycles. It is assumed a district of area $A = 40$ sq.km. located 7 km from the depot. The expected driving line-haul time from the depot to the district is $t_{LH1} = 14$ min and the line-haul time from the last service to the depot is admitted to be the same, i.e. $t_{LH2} = t_{LH1} = 14$ min, with a standard deviation $\sigma_{LH2} = \sigma_{LH1} = 2.8$ min, being both normally distributed. The servicing time $t_j^{(ST)}$ at a generic client location, from the instant the vehicle stops until it leaves to attend another client, is assumed to be described by a lognormal distribution, with constant $E[t_j^{(ST)}] = 11$ min, and $\sigma_j^{(ST)} = 4.5$ min. Assuming n servicing points randomly generated over the district, a combination of farthest insertion and *3-OPT* algorithms (Syslo et al. 2006) was applied in order to get the Travelling Salesman Problem (*TSP*) Euclidean route. A corrective coefficient (route factor) was then applied to the Euclidean distance to take into consideration the road network impedance (Novaes and Burin 2009).

The vehicle average speed within the district under standard traffic conditions, is $s_0 = 28$ km/h, with the velocity reduced to $s_1 = 15$ km/h during over-congested situations. The speed, both in a normal traffic condition and during an over-congested situation, is represented by lognormal distributions (section Sequential Analysis for Detecting Incipient Faults). It is assumed that over-congested situations occur with probability $p_1 = 0.20$ (hypothesis H_1), while standard conditions prevail with probability $p_0 = 0.80$ (hypothesis H_0).

Let $H = 8$ h be the maximum vehicle crew working time per day and T_C the vehicle cycle time, with $T_C \leq H$. Assuming a number n of servicing points in the district route, one is interested in estimating the expected number of visits that will be effectively performed during the daily cycle-time, with the objective of keeping it within a pre-established service level. Since n is sufficiently large in our applications, T_C can be assumed to be normally distributed according to the central limit theorem. Thus, it is necessary to estimate an upper extreme value for T_C in order to select an appropriate value for n.

Statistics of extremes have applications in many engineering domains (Gumbel 1967; Smith 2003; Haan and Ferreira 2006). Consider M samples of an i.i.d. continuous random variable X, each sample of size m taken from the same population. The asymptotic configuration, provided it exists, must be such that the largest value of any sample of size m taken from the population must have the same distribution (Gumble 1967). Let $\{X_1, X_2, \ldots, X_m\}$ represent one such sample. Let $Y_m = \max(X_1, \ldots, X_m)$ be the sample maximum. The probability that the largest value is below a generic value x is

$$Pr\{Y_m \leq x\} = F(x)^m, \tag{1}$$

$F(x)$ being the cumulative probability distribution function of x. Since a linear transformation does not change the form of the probability distribution, the

probability that the largest value is below x is equal to the probability of a linear function of x (Gumbel 1967)

$$F(x)^m = F(a_m x + b_m), \tag{2}$$

the two parameters a_m and b_m being functions of m. It has been proved that, as $m \to \infty$, expression (2) tends to the cumulative probability distribution $H(x)$ of the Gumbel type in case X is described by exponential, normal, lognormal, or gamma distributions. The Gumbel distribution is (Gumbel 1967; Smith 2003)

$$\lim_{m\to\infty} F(x)^m = H(x) = exp\{-exp[-\alpha(x-u)]\}, \tag{3}$$

where α and u are coefficients obtained via calibration. Let $\hat{E}[X]$ and $\hat{\sigma}[X]$ be, respectively, the expected value and the standard variation of a sample formed by the upper extreme values extracted from M sets containing, each set, m values of a continuous variable X. Then, the estimated values of α and u are (Gumbel 1967)

$$\frac{1}{\hat{\alpha}} = \frac{\sqrt{6}}{\pi}\hat{\sigma}[X], \text{ and} \tag{4}$$

$$\hat{\sigma}\,\hat{u} = \frac{\pi}{C_V\sqrt{6}} - \gamma, \tag{5}$$

where $C_V = \widehat{\sigma}[X]/\widehat{E}[X]$ is the coefficient of variation and $\gamma = 0.57722$. From (4) and (5) one gets

$$\widehat{u} = \widehat{E}[X] - \frac{0.57722}{\hat{\alpha}}. \tag{6}$$

We are interested in determining the maximum value of the vehicle cycle time $T_C^{(\max)}$. Letting $x = T_C^{(\max)}$, making $H(T_C) = \vartheta$ in (3), where ϑ is a pre-assumed confidence level, and simplifying

$$T_C^{(\max)} = \widehat{u} - \frac{\ln[-\ln(\vartheta)]}{\hat{\alpha}}. \tag{7}$$

A simulated data set was generated consisting of daily cycle times forming $M = 15$ blocks, each block containing $m = 30$ simulated sample values of T_C, representing a total of 450 samples for each value of n. To perform the simulation it is necessary to assume a value for n beforehand. The objective is to get a maximum value of n such that $T_C \leq H$. Five values of n were tested, as shown in Table 1. For a specific value of n, fifteen simulated blocks were produced, yielding 15 values of $T_C^{(\max)}$, one for each block. The average and standard deviation of $T_C^{(\max)}$ were obtained as shown in Table 1. Expressions (4) and (6) yielded the values of $\hat{\alpha}$ and $\hat{u}$. Assuming a $\vartheta = 0.98$ confidence level, expression (7) furnishes the overall maximum $\widehat{T}_C^{(\max)}$. We selected $n = 22$, with a value of $\widehat{T}_C^{(\max)}$ close to the permitted limit of 8 h.

Table 1 Searching for n such that $T_C \leq H$

N	$\hat{E}[T_C]$	$\hat{\sigma}[T_C]$	$\hat{\alpha}$	$\hat{u}$	$\hat{T}_C^{(\max)}$(h)
20	7.12	0.2122	6.0443	7.0285	7.58
21	7.40	0.1987	6.4547	7.3086	7.84
22	7.61	0.2100	6.1074	7.5208	8.04
23	7.96	0.1983	6.4677	7.8687	8.36
24	8.14	0.2117	6.0589	8.0480	8.53

Table 2 shows the static vehicle tour simulation framework. The line-haul travelling times from the depot to the district (outbound) and vice versa (inbound) are represented by t_{LH1} and t_{LH2} respectively in Table 2. The occurrence of hypothesis H_0 or H_1 is represented by h. Variables $d_{i-1,i}$ and $t_{i-1,i}^{(h)}$ are the distance and the travelled time, respectively, from point $i-1$ to point i in the route. The vehicle speed s is represented by log-normal distributions, which depend on the occurrence of H_0 or H_1 (section Sequential Analysis for Detecting Incipient Faults). The stopping time at point i is $t_i^{(ST)}$, and τ_i is the cumulative elapsed time up to stage i.

Table 2 Static route simulation framework

1. Assume a value for n and compute point density $\delta = n/A$;
2. $i \leftarrow 0$;
3. Generate line-haul travelling times, t_{LH1} and t_{LH2}, both normally distributed;
4. Set $\tau_0 \leftarrow t_{LH1} + t_{LH2}$;
5. Generate random number ε; if $\varepsilon \leq p_0$, $h \leftarrow 0$; if $\varepsilon > p_0$, $h \leftarrow 1$ (*);
6. $i \leftarrow i + 1$ (i is the servicing point sequencial number);
7. Generate value for $d_{i-1,i}$, Erlang distributed with parameter $\theta = 3$;
8. Generate value for the route factor k_2 (log-normal);
9. Generate stopping time $t_i^{(ST)}$ at point i, log-normally distributed;
10. If $h = 0$, then $E[s] \leftarrow s_0$, else $E[s] \leftarrow s_1$.Generate speed value s (log-normal)
11. $t_{i-1,i} \leftarrow d_{i-1,i} \times k_2/s$;
12. $\tau_i \leftarrow \tau_{i-1} + t_{i-1,i} + t_i^{(ST)}$
    (a) If $\tau_i \leq H$ and $i < n$, then $n^{(P)} \leftarrow n$ and $T_C = \tau_i$; go to (6);
    (b) If $\tau_i \leq H$ and $i = n$, then
    Begin
    $n^{(P)} \leftarrow n$; $T_C \leftarrow \tau_i$; go to (14);
    End
    (c) If $\tau_i > H$ then
    Begin
    $n^{(P)} \leftarrow [n - (i-1)]$; $T_C \leftarrow \tau_{i-1}$; go to (14);
    End
13. $n^{(U)} \leftarrow [n - n^{(P)}]$
14. Repeat the process from (2) on until the number of replications is complete

(*) $h = 0$, hypothesis H_0 and $h = 1$, hypothesis H_1;

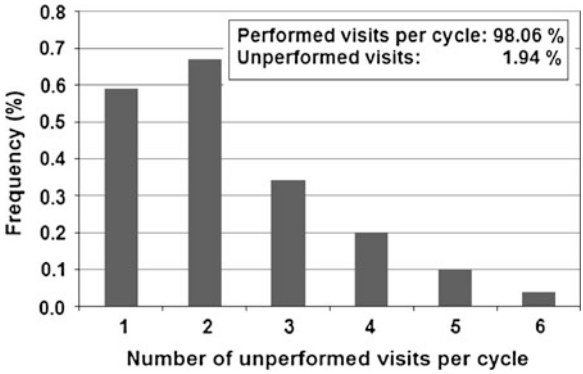

Fig. 1 Static routing: unperformed tasks per vehicle cycle

The vehicle cycle time is T_C; the number of performed tasks in the cycle is $n^{(P)}$, and the number of unperformed tasks is represented by $n^{(U)}$.

Running the simulation model for $n = 22$, with 20,000 replications and assuming over-congested traffic conditions in 20 % of the working days, it led to a 1.94 % rate of unperformed tasks during a typical working cycle. Figure 1 shows the distribution of unperformed tasks per cycle. Most frequent situations are the ones with 1 or 2 unperformed visits per cycle (1.26 % of de cases from a total of 1.94 %). On the other hand, considering only the days when over-congested traffic conditions occur, the average rate of unperformed visits raised to 10.5 %, more than five times the former value, leading to a service level much lower than the desired value.

This result deserves some considerations. First, certain traffic disruptions have severe impacts on vehicle movement as, for example, the ones caused by public transport strikes, heavy rains, etc., and they may last for some days. Those situations, when covering somewhat longer periods, may generate excessive back-logs of tasks, thus impairing the logistics operations for some time. Second, since JIT operations are quite common in global supply chains, unpredictable delays, as the mentioned situations, will lead, in the long term, to additional safety stock compensations in order to maintain the manufacturer's production line uninterrupted. With these aspects in mind, it is apparent that some improving measures are opportune.

The Dynamic Routing Problem

To avoid unexpected backlog in the pick-up process described in section The Static Routing Problem, a simple alternative is available to the logistics operator, although potentially costly: reduce the number of visiting points per vehicle by putting more trucks to perform the service. With this measure, the risk of unperformed tasks obviously will be reduced. But, as a result of the great number of random components that form the vehicle cycle time, its average value will be

relatively small, meaning a low fleet usage rate. Another operating alternative is to establish a cooperative scheme, where part of the planned tasks assigned to a truck can be transferred to auxiliary vehicles whenever an excessive service load is foreseen by the on-board computer system.

Such a scheme is part of a new form of managing integrated logistic services in the supply chain, in which a set of intelligent agents, each responsible for one or more activities and interacting with other agents in planning and performing their tasks. In this new form of acting, an agent is an autonomous goal-oriented software process that operates asynchronously, communicating and coordinating with other participant agents as needed (Fox et al. 2000; Davidsson et al. 2005; Berger and Bierwirth 2010).

The smallest controlling entity in this approach (an agent) is described as anything that is able to "perceive its environment through sensors (hardware and software) and act upon that environment through actuators" (Russel and Norvig 2003). A Multi-Agent System (MAS) is a system consisting of independent intelligent control units linked to physical or functional entities such as vehicles, orders, etc. (Mes et al. 2007). Agents act autonomously by pursuing their own objectives and interact with each other using informational exchange and negotiation mechanisms (Mes et al. 2007). In this application the agents are the vehicles which perform the on-route tasks, plus the central depot which has supplementary vehicles that can be eventually assigned to the routes in case other agents do not reach agreement to exchange tasks. The objective is to eliminate or reduce as much as possible the number of unperformed tasks in the pick-up vehicle routing problem described in section The Static Routing Problem.

Fault detection and fault diagnosis methods will be employed in this work to dynamically anticipate operational counteractions in order to avoid unexpected unperformed tasks along the route. A fault-detection software is to be installed aboard, and the vehicle is assumed to be provided with a geo-referencing device and telecommunication equipment linking the vehicle to nearby collaborative agents (other vehicles and the central depot). The rationale involved will permit to infer, during the servicing process, if traffic conditions will impair the accomplishment of the planned tasks during a working day, thus transferring part of the jobs to other vehicles (agents), and leading to a collaborative scheme among them.

Fault Detection and Fault Diagnosis

The concepts and definitions of fault detection and diagnosis set forth in this section are based mostly on Isermann (1997, 2005). Other references are Basseville and Nikiforov (1993), Haan and Ferreira (2006) and Simani et al. (2010). A fault is defined as an unpermitted deviation of at least one characteristic property of a variable from an acceptable behaviour. Therefore, the fault is a state that may lead to a malfunction or failure of the system. The time dependency of faults can be classified as (a) *abrupt fault* (stepwise), (b) *incipient fault* (drift like), and (c) *intermittent*

fault. One calls *abrupt fault* any change in the parameters of a system that occurs either instantaneously or at least very fast with respect to the sampling period of the measurements. Abrupt faults do not refer to changes in the process with large magnitude; in fact, in most applications the main question is how to detect small changes. On the other hand, incipient faults are also of interest: they are behaviour deviations that occur cumulatively over time, yet not failures, but leading to future underperformance or disruptions in the system. In our application we will deal with detection of incipient and abrupt faults.

To control technical systems, supervisory functions are installed to indicate undesired or unpermitted process states, as well as to take appropriate actions in order to maintain the operation within a pre-established service level and to avoid unexpected disruptions. Three basic function types can be distinguished (Isermann 1997): (a) *monitoring*, in which measurable variables are checked with regard to tolerances and warnings are generated for the operator; (b) *automatic protection*, normally used to control dangerous processes that cannot wait for external interventions; and (c) *supervision with fault diagnosis* which is based on the measurement of some key variables and the calculation of parameters, resulting in the identification of symptoms via change detection, followed by fault diagnosis, and leading to counteraction decisions. Type (c) function is compatible with in-depth fault diagnosis, either with *abrupt* or *incipient* time behaviour, being more comprehensive. In our application supervisory functions of type (a) and (c) will be employed.

The goal for early fault detection and diagnosis is to have enough time for counter-actions such as adding supplementary operations, reconfiguration, maintenance or repair, etc. The earlier detection can be achieved by gathering more information, especially by using the relationship among the measurable quantities in the form of mathematical models. For fault diagnosis, the knowledge of cause-effect relationships has to be used. In our analysis, two types of faults will be considered: (a) *incipient faults*, occasioned by over-congested traffic conditions, in which the measurement and analysis of commanding parameters occur cumulatively over time, and (b) *abrupt faults*, represented by unpredictable delays that may occur at *OEM* premises when transferring goods to the logistics operator, as well as vehicle breakdowns during the cycle, or another sort of exceptional random interruptions.

Incipient Faults Occasioned by Exceptional Traffic Congestion

Traffic congestion is seen as a condition of traffic delay (i.e., when vehicle flow is slowed below reasonable speeds) because the number of vehicles trying to use a road exceeds the capacity of the network to handle it (Weisbrod et al. 2003). In addition to speed reduction, congestion also introduces variability in traffic conditions, which is known as *travel time reliability* (Cambridge Systematics 2005). The resulting traffic slowdowns and travel time reliability produce negative effects

on supply chain activities, including impacts on vehicle traveling costs, air quality and noise, labour efficiency, industrial and commercial productivity, customer service level, etc. The severity and pattern of congestion, as well as the effectiveness of alternative policies and interventions to address it, vary widely from place to place. That can depend on the size and layout of the urban area, its available transportation options, and the nature of traffic generators (Weisbrod et al. 2003). Congestion is usually the result of seven root causes, often interacting with one another (Cambridge Systematics 2005):

1. Physical bottlenecks, such as reduced number of lanes, narrow lane and shoulder widths, inadequate roadway grades and curves, etc., leading to reduced road capacity.
2. Traffic incidents, occasioned by events that disrupt the normal flow of traffic, such as vehicular crashes, breakdowns, etc.
3. Work zones temporally reserved for construction and repair activities on the roadway, generating lane reduction, narrower traffic spaces, lane shifts, detours, speed reduction, etc.
4. Weather conditions such as snow, flood, fog, etc., that can lead to substantial changes in driver behaviour,
5. Traffic control devices, which leads to intermittent disruption of traffic flow, such as railroad grade crossings, poorly timed light signals, traffic interferences with street cars and bus ways, etc.
6. Special events that cause severe traffic flow variations in its vicinity, being radically different from typical day-to-day patterns. For instance, a rapid transit labor strike in a city with heavy public transport patronage and generating additional flow of cars and busses.
7. Fluctuations in normal traffic which shows, on a day-to-day basis, fluctuations with days with higher traffic volumes than others. In cities with constant heavy traffic volumes, even random fluctuations can result in unreliable and over-congested traffic conditions.

Another important traffic congestion classification is due to Brownfield et al. (2003). The first type is *recurrent congestion*, which can be anticipated by road users that are acquainted with the route. The other type is *non-recurrent congestion*, which occurs at non-regular times at a site. It is unexpected and unpredictable to the driver. In our analysis, it is assumed that the logistics entity in charge of the urban transport service is aware of all programmed events, i.e. it is fully prepared to cope with recurrent congestion. Thus, from the seven factors listed above, causes (1), (3) and (5) are not considered in our application. Conversely, it is assumed that over-congested situations are originated by causes (2), (4), (6) or (7).

Although there is no existing, universally accepted, quantitative definition for traffic congestion, its analysis must rely on easy to measure elements if its impacts are to be evaluated and compared across the range of situations considered in the investigation (Brownfield et al. 2003). One frequent assumption is to assume that an urban road link is congested if its average speed is below a given upper threshold. In addition to average speed reduction to travellers, the sources of congestion also

produce time variability known as *travel time reliability* (Cambridge Systematics 2005), which can be defined in terms of how travel times vary within a pre-defined period. In practical terms, it is useful to fit statistical frequency distributions to travel time, to see how much variability exists in critical sites of the road network.

Exceptionally, unpredictable and heavy traffic congestions caused by severe accidents, public transport strikes, heavy storms, etc., may occur during certain working days. In these situations the travelling speed decreases sharply. Let s_0 be the average travelling speed in a route in a generic working day, and suppose the average speed reduces to a level $s_1 \ll s_0$ when over-congested situations occur. Then, one could say that the traffic conditions are normal if $s \geq s_0$, and over-congested if $s \leq s_1$. Moreover, if $s_1 < s < s_0$, one would not decide immediately for either alternative, waiting for more information to take a decision. Of course, this is a typical statistical hypothesis testing. Nevertheless, an instantaneous travelling time increase is not, in itself, an indication of an over congested situation. In fact, many non-recurrent events have short duration, and their effects dissipate more or less rapidly. Furthermore, some recurrent events have local impact only, and their effects do not extend to other parts of the served region. Over-congested situations that are of interest in our analysis are the ones with broader geographical extension and longer duration, although in many cases they are no longer than 24 h. Thus, travel time reliability covering an expressive subset of the urban region, seems to be a good judgmental criterion to evaluate it. And, in order to measure travel time reliability it is necessary to sequentially collect and analyse traffic data. With today's on-board telematics and computing devices it is not difficult to collect and analyse real-time information on travelled distance, time and speed with satisfactory accuracy (Goel 2008). In our study, the statistical inference process to detect an over-congested condition follows a sequential analysis methodology (Wald 1947; Basseville and Nikiforov 1993; Lai 2001), which is described in the next section.

Dynamic Detection of Over-Congested Traffic Conditions

Day-to-day traffic flow variability in urban networks produces typical traffic patterns, but unexpected events cause occasional surges in traffic volumes that overwhelm the road system. Such events, of a "hectic" pattern, are generated by accidents with severe traffic interruptions, extensive public transport strikes, and long duration storms, among others. Strong changes in some characteristic properties of a system may occur occasionally in both technological and natural environments. And due to today's availability of information processing systems, complex monitoring algorithms have been developed and implemented (Basseville and Nikiforov 1993). The key difficulty in detecting a fault occurrence through the observation of some properties of a system is to separate noise from the relevant factors. In addition, some failures have a catastrophic nature, leading to an abrupt change in the control variables. But some faults occur with gradual changes in the

system attributes over time. One way of tackling the latter is Sequential Analysis (Wald 1947; Basseville and Nikiforov 1993; Lai 2001).

Classical techniques of statistical inference and hypotheses testing adopt a fixed sample size. With this kind of approach one seeks to minimize the error probabilities for a given sample size. The size of the sample is defined beforehand, and following its statistical analysis one of two possible actions is taken: accept the null hypothesis H_0, or accept the alternative hypothesis H_1. The null hypothesis represents in our analysis the standard or basic situation, whereas the alternative hypothesis indicates the occurrence of an abnormal condition, leading to a fault in the system. Another way to solve hypotheses testing problems, when the sample size is not fixed a priori but depends upon the data that have already been observed, is Sequential Analysis. Now the problem is: for given error probabilities, try to minimize the sample size, or equivalently, make the decision with as few observations as possible. Contrary to the fixed sample size approach, a third possible course of action may occur in sequential analysis when the evidence is ambiguous: take more observations until the evidence strongly favours one of the two hypotheses. Thus, sequential analysis follows a dynamic sequence of observations in such a way that the decision to terminate or not the experiment depends, at each stage, on the previous test results.

Although some authors date the rudiments of sequential analysis to the works of Huyghens, Bernoulli, and Laplace, such methodology was effectively born in response to demands for more efficient testing of anti-aircraft gunnery during World War II, culminating with the development of the *Sequential Probability Ratio Test* (*SPRT*) by Wald, in 1943 (Lai 2001). A typical case of sequential estimation arises when only two unknown parameters μ and σ are required to define the distribution of the random variable x object of our analysis. Let $f(x, \mu, \sigma)$ denote the probability density function of x, when x is continuous. Conversely, if x is discrete, $f(x, \mu, \sigma)$ represents its probability. Let $x_1, x_2, \ldots, x_m$ be a set of m sequential and independent observations on x. Due to the independence of the observations, the joint probability density function is

$$f(x_1, \mu, \sigma) f(x_2, \mu, \sigma) \ldots f(x_m, \mu, \sigma). \tag{8}$$

Suppose that the distribution of the random variable x under consideration is defined by q unknown parameters (in our case, $q = 2$). A statement about the values of the q parameters is called a *simple hypothesis* if it determines uniquely the values of all q parameters. It is called a *composite hypothesis* if it is consistent with more than one value for some parameter (Wald 1947). Let us analyse the test of simple hypothesis that $\mu = \mu_0$ and $\sigma = \sigma_0$, where μ and σ are the expected value and the standard deviation of the probability distribution of x. This hypothesis is the null hypothesis denoted by H_0. The alternative hypothesis that $\mu = \mu_1$ and $\sigma = \sigma_1$ will be denoted by H_1. Thus, we shall deal with the problem of testing the simple hypothesis H_0 against the alternative simple hypothesis H_1, on the basis of a sample of m independent observations $x_1, x_2, \ldots, x_m$ on x. According to the developments of Neyman and Pearson, errors of two kinds are present when one

accepts or rejects hypothesis H_0. We commit an error of first kind if we reject H_0 when it is true. On the other hand, we commit an error of the second kind if we accept H_0 when $H_1$is true. We denote the probability of an error of the first kind by α, and the probability of an error of second kind by β.

To apply the *SPRT* developed by Wald (1947) for testing $H_0 : \mu = \mu_0, \sigma = \sigma_0$ against $H_1 : \mu = \mu_1, \sigma = \sigma_1$, two positive constants A and B ($B < A$) are computed

$$A = (1 - \beta)/\alpha \text{ and } B = \beta/(1 - \alpha) \tag{9}$$

Suppose one has drawn m samples leading to the independent observations $x_1, x_2, \ldots, x_m$ on the random variable x. At this stage of the experiment the *SPRT* (Wald 1947; Basseville and Nikiforov 1993; Lai 2001) is computed as

$$\pi_m = \frac{f(x_1, \mu_1, \sigma_1) f(x_2, \mu_1, \sigma_1) \ldots f(x_m, \mu_1, \sigma_1)}{f(x_1, \mu_0, \sigma_0) f(x_2, \mu_0, \sigma_0) \ldots f(x_m, \mu_0, \sigma_0)}. \tag{10}$$

Three situations may occur:

1. If $B < \pi_m < A$, the experiment continues by taking an additional observation;
2. If $\pi_m \geq A$, the experiment terminates with the rejection of H_0;
3. If $\pi_m \leq B$, the experiment terminates with the acceptance of H_0.

For purposes of mathematical simplification, it is more convenient to compute the logarithm of the ratio π_m. Let

$$z_i = ln\left(\frac{f(x_i, \mu_1, \sigma_1)}{f(x_i, \mu_0, \sigma_0)}\right). \tag{11}$$

Define

$$\pi_m^* = \ln(\pi_m) = z_1 + z_2 + \cdots + z_m. \tag{12}$$

The test is addictive now. The experiment continues if $\ln B < \pi_m^* < lnA$ by taking an additional observation; the process terminates with the rejection of H_0 *if* $\pi_m * \geq lnA$; and it terminates with the acceptance of H_0 if $\pi_m * \leq lnB$.

In practical cases, composite hypothesis may occur. One way to solve sequential analysis problems with composite hypothesis is the method of a weighting function associated with the generalized likelihood ratio algorithm (Basseville and Nikiforov 1993). To do this, two weighting probability distributions, with density functions $g(H_0)$ and $g(H_1)$, depending on H_0 and H_1 respectively, are introduced into the model. The *SPRT* is now transformed into a weighted likelihood ratio test (Basseville and Nikiforov 1993). But, in order to do this, it is necessary to fit distributions $g(H_0)$ and $g(H_1)$ to the data, which depends on detailed information not commonly available in real settings. In the application considered in this paper, a more tractable composite hypothesis test is adopted. This composite hypothesis testing is represented by $H_0^{'} : \mu \leq \mu_0, \sigma \leq \sigma_0$ versus $H_1^{'} : \mu \geq \mu_1, \sigma \geq \sigma_1$, such that $\mu_1 > \mu_0$ and $\sigma_1 > \sigma_0$. This model is usually sufficient for practical purposes (Lai 2001). Assuming that the probabilities of the

errors of first and second kind also do not exceed α and β, one can use the *SPRT* of the simple hypothesis $H_0 : \mu = \mu_0, \sigma = \sigma_0$ versus $H_1 : \mu = \mu_1, \sigma = \sigma_1$, with the same error probabilities α and β. However, while this *SPRT* has minimum expected sample size at $\mu = \mu_0$, $\sigma = \sigma_0$ and at $\mu = \mu_1, \sigma = \sigma_1$, its maximum expected sample size over μ and σ can be larger than the optimal fixed sample size (Lai 2001). This means that sometimes the sequential test will not be sufficient to detect hypothesis H_1 during the daily tour, generating unperformed tasks at the end of the working day. But unperformed tasks will be eliminated or drastically reduced when compared with the static alternative, as it will be shown in section Sequential Analysis for Detecting Incipient Faults, a fact that justifies the adoption of the dynamic setting in our model.

Sequential Analysis for Detecting Incipient Faults

In this application, the variable that commands the decision whether to seek help from another agent or to proceed along the planned routing process is the vehicle speed s. In fact, since link lengths vary along the route, and consequently the resulting displacement times also vary, speed is a more appropriate variable to measure traffic variations. The renewal epoch (stochastic regenerating point) of the sequential decision process is defined as the instant when the vehicle crew has just terminated a task at an OEM location and is ready to depart for the next visit. At such an instant, the on-board computer evaluates the displacement time $t_{i-1,i}$ over the traveled segment linking the last visiting stop $i-1$ to the present one i. The corresponding speed s is simply obtained by dividing the travelled segment extension by its respective displacement time, both elements assumed to be available on the on-board system. For the district under analysis it is assumed that there are enough historical data on speed values covering the standard traffic condition and the over-congested scenario. In particular, the average speed $s \geq s_0$ is related to standard traffic conditions and $s \leq s_1$ represents the over-congested scenario. This information, together with the series of data collected up to that point, will serve as the basis for inferring whether the traffic is normally behaved or is over-congested, thus leading to the appropriate operational decision.

As discussed in section Dynamic Detection of Over-Congested Traffic Conditions, it is necessary to define a probability distribution $f(x, \mu, \sigma)$ to represent the random variable that commands the decision process. A sample was gathered in a representative route located in the urban area under analysis, involving 40 vehicle travel speeds during typical working days. A log-normal distribution was fitted to the data (Fig. 1):

$$f(s) = \frac{1}{\sigma\sqrt{2\pi}} exp\left\{ -\frac{1}{2}\left[\frac{\ln(s) - \mu}{\sigma}\right]^2 \right\}, \; s > 0, \tag{13}$$

where s is the speed in km/h, and μ and σ are the parameters of the log-normal distribution given by

$$\mu = ln\left\{\frac{E[s]^2}{\sqrt{var[s] + E[s]^2}}\right\} \text{ and } \sigma = \sqrt{ln\left\{\frac{var[s]}{E[s]^2} + 1\right\}}, \tag{14}$$

where $E[s]$ and $var[s]$ are the expected value and the variance of s respectively. For the mentioned sample of travelling times, assumed to represent the normal traffic conditions in our application, one has $E[s] = 28$ km/h and $var[s] = 44.84$, leading to $\mu_0 = 3.3044$ and $\sigma_0 = 0.2354$.

For the over-congested traffic conditions one has $E[s_1] = 15$ km/h. It was assumed that, for this situation, the speed is also represented by a log-normal distribution. It was assumed further that the coefficient of variation is the same for hypotheses H_1 and H_0, i.e. $C_V = \sqrt{44.84}/28.0 = 0.239$. Depending on real data, this assumption may be changed, fitting a value of $var[s_1]$ directly over the real data. Thus, for the over-congested traffic condition in our application one has $var[s_1] = C_V^2 \times E[s_1]^2 = 12.852$, leading to $\mu_1 = 2.680$ and $\sigma_1 = \sigma_0 = 0.2354$. Substituting (13) into (11) and (12), and making the necessary simplifications, one gets the *SPRT* parameter,

$$\pi_m^* = m\, ln\left(\frac{\sigma_0}{\sigma_1}\right) + \frac{1}{2}\sum_{i=1}^{m}\left[\frac{\ln(s^{(i)}) - \mu_0}{\sigma_0}\right]^2 - \frac{1}{2}\sum_{i=1}^{m}\left[\frac{\ln(s^{(i)}) - \mu_1}{\sigma_1}\right]^2, \tag{15}$$

where m is the sequential number of the test and $s^{(i)}$ is the vehicle speed measured at stage i along the route within the district (Fig. 2).

At each regeneration point (stage) the *SPRT* value π_m^* (15) is computed. Depending on the *SPRT* value, three scenarios are defined:

(a) scenario $sc = 0$, when $\pi_m^* \leq B$;
(b) scenario $sc = 1$, when $\pi_m^* \geq A$;
(c) scenario $sc = 2$,when $B < \pi_m^* < A$.

Countermeasures are taken if scenario $sc = 1$ occurs; otherwise, the routing process continues unchanged until the next stage. Figure 3 shows a schematic representation of the vehicle routing sequence and the decision stage where the *SPRT* is performed. Assume that the vehicle agent AG_A left the depot with the assignment of n visits. Suppose the sequential test indicates the occurrence of hypothesis H_1 at stage 3, as shown in Fig. 3. At that point, the on-board computer checks how many of the remaining visits should be transferred to another vehicle agent. Let k be the number of visits to be transferred. Upon negotiation, agent AG_B agrees to perform the k tasks. Of course, depending on the number of visits to be transferred, more than one agent can be involved in the transference.

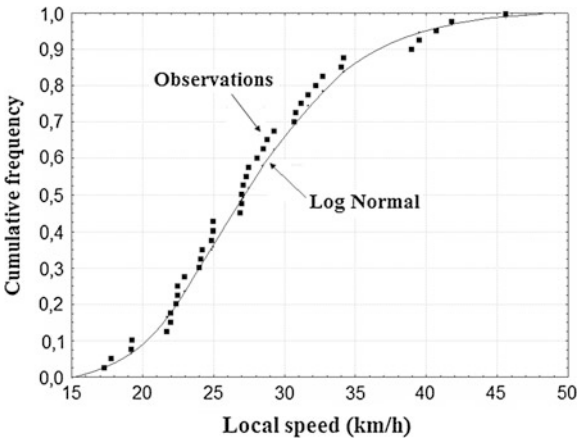

Fig. 2 Fitting a log-normal distribution to the local speed

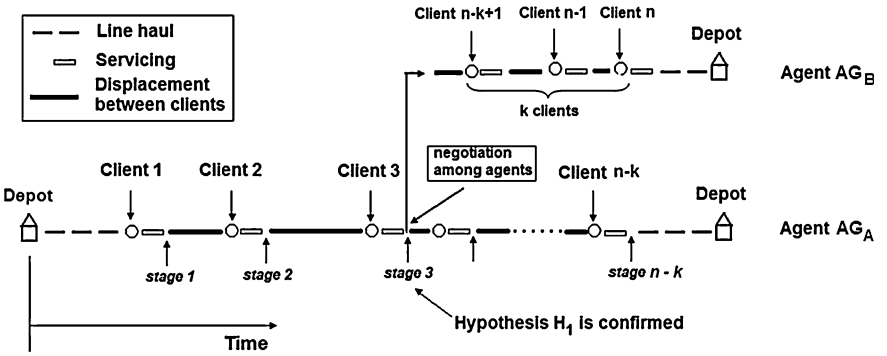

Fig. 3 The vehicle routing sequence and the decision stage

Transference of Tasks

Let i be the actual stage of the routing process. Recall that stage i corresponds to the instant when the ith pick-up service has just terminated, and the vehicle is ready to depart to the next stop. Let us analyse first the occurrence *of incipient faults*. Define $\tau_i^{(h)}$ as the cumulative vehicle time along the route, measured from the departure from the depot, up to stage i, given hypothesis h (either $h = 0$, or $h = 1$) occurs. It is given by

$$\tau_i^{(h)} = t_{LH1} + \sum_{j=2}^{i} t_{j-1,j}^{(h)} + \sum_{j=1}^{i} t_j^{(ST)}. \tag{16}$$

On the other hand, let $\omega_{i,j}^{(h)}$ be the elapsed time from the actual stage i to stage j ($j > i$), assuming hypothesis h occurs, and also assuming that the vehicle returns to the depot just after visit j

$$\omega_{i,j}^{(h)} = \sum_{m=i+1}^{j} t_{m-1,m}^{(h)} + \sum_{m=i}^{j} t_m^{(ST)} + t_{LH2} \tag{17}$$

We have assumed in the application that all random variables are independent. Due to the central limit theorem and for i sufficiently large, variable $\omega_{i,j}^{(h)}$ can be approximately represented by a normal distribution. Thus, for a 98 % significance level, the maximum expected value of $\omega_{i,j}^{(h)}$ is

$$\overline{\omega}_{i,j}^{(h)} = \max \omega_{i,j}^{(h)} \cong E[\omega_{i,j}^{(h)}] + 2.06\sqrt{var[\omega_{i,j}^{(h)}}} \tag{18}$$

At the beginning of the routing process is not yet known which traffic condition is prevailing, thus hypothesis H_0 (which shows higher probability) is assumed, i.e. $h = 0$. Further, at every stage i the SPRT is performed. Suppose scenario $h = 0$ occurs. This indicates that the routing process should proceed unchanged, with hypothesis H_0 prevailing. On the other hand, suppose scenario $h = 1$ occurs, meaning one should accept hypothesis H_1. Then, in order to define the possible number of visits to be performed in the route, one has to seek for the largest value of j such that the expected total time to accomplish the tasks is not greater than H

$$\tau_i^{(h)} + \bar{\omega}_{i,j}^{(h)} \leq H, \text{ with } h = 1 \text{ and } j > i. \tag{19}$$

Thus, the total number of visits to be performed in the tour is $n' = i + j$, and the number of visits to be transferred to other agents is $n^T = n - (i + j)$, which represents the expected number of *incipient faults* in the application. Of course, since n^T tasks are transferred, there will be less tasks remaining to be considered at the next stages of the process, i.e. $n \leftarrow n'$, with $n' < n$.

After handling incipient faults, the model investigates the occurrence of *abrupt faults*. Here, an abrupt fault refers to the occurrence of unperformed tasks at the end of a daily vehicle cycle occasioned by exceptional unpredictable delays at *OEM* premises when picking-up manufactured orders, lorry breakdowns when travelling along the route, etc. At each decision stage the on-board system estimates the maximum cycle time to perform all visits, considering the effective elapsed time so far, plus the eventual observed delays and the remaining visits to be done. If the daily cycle time limit is surpassed, the on-board system estimates how many visits are to be transferred to other agents.

Suppose an exceptional and unpredictable manufacturing delay Δ_i occurs at stage i. Care must be taken not to consider as exceptional delays situations already contemplated in historical data variability. An exceptional delay ∇_m may also occur at a transport link m. Relation (19) is now modified as follows

$$\tau_i^{(h)} + \overline{\omega}_{i,j}^{(h)} + \Delta_i + \nabla_m \leq H, \text{ with } h = 0 \text{ or } 1. \quad (20)$$

Again, one looks for the largest value of j such that the cycle time constraint is respected, and estimating the number of remaining visits that will be performed and the tasks that have to be transferred to other agents.

Simulation Results

The simulation of the dynamic model for $n = 22$, with 20,000 replications and assuming over-congested traffic conditions in 20 % of the working days, resulted in a 1.63 % rate of transferred tasks during a typical working cycle. From that, 1.50 % of the transferences were generated by incipient faults, and 0.13 % by abrupt faults. The transference total of 1.63 % is less than the 1.94 % level of unperformed tasks observed in the static case. This happens because, as the sequential test is performed, followed by abrupt fault detection, the number of remaining tasks to be done decreases, thus reducing the possibility of occurring other additional faults. Figure 4 shows that the occurrence of four transferred tasks per cycle is the most frequent situation in the dynamic case. In fact, since it takes some time until the *SPRT* can detect hypothesis H_1, the number of prospective tasks to be transferred tends to increase, and are likely to happen all at the same moment. Abrupt tasks also occur, but at a significantly reduced frequency.

On the other hand, considering only the days when hypothesis H_1 occurred, the rate of transferred tasks was 7.93 %, from which 7.50 % was generated by incipient faults and 0.43 % by abrupt faults. In those two situations it was admitted $\Delta_i = 0$ and $\nabla_m = 0$ in (20), meaning that abrupt faults were not generated by exceptional delays, but were occasioned by intrinsic variations in the random variables that form the cycle time. Of course, if those elements were not nil, the results would reflect their presence.

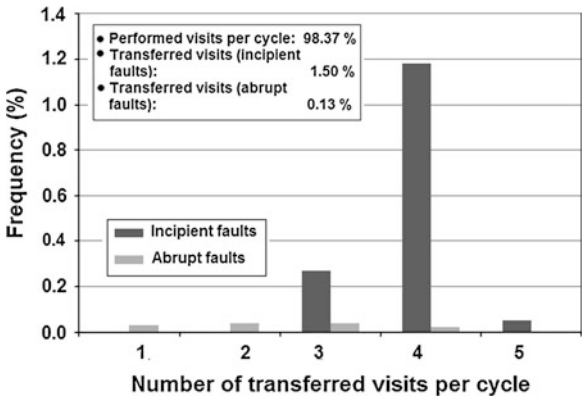

Fig. 4 Dynamic routing: transferred tasks per vehicle cycle

Conclusions

Although potentially suitable for applications, a number of points are still open for further research. First, it remains to be investigated the criteria to decide which tasks should be transferred to other agents, considering the vehicle routing configuration and the corresponding schemes of other prospective agents. Another important point is that, depending on the *OEM* plant locations and the time of the day, it might be impossible to transfer tasks, requiring other anticipating measures from the central depot. A third question refers to the routing optimization process based on the *TSP* criterion, which is the prevalent case in most dynamic routing problems reported in the literature, where one searches for the route that minimizes travelled distance or time. If the components or products of the diverse *OEM* manufacturing plants show different added values, the ones with the highest values should not be located at the end of the picking up process due to the higher inventory costs occasioned by unperformed tasks. Further research is planned with the objective of developing new heuristics to solve this kind of vehicle routing problem incorporating product value considerations in the optimization criteria.

Acknowledgments This research has been supported by the Brazilian Capes Foundation and by DFG — German Research Foundation, Bragecrim Project n° 2009-2.

References

Basseville M, Nikiforov I (1993) Detection of abrupt changes: theory and application. Prentice-Hall, New Jersey

Berger S, Bierwirth C (2010) Solutions to the request reassignment problem in collaborative carrier networks. Transp Res Part E 46:627–638

Brownfield J, Graham A, Eveleigh H, Maunsell F, Ward H, Robertson S, Allsop R (2003) Congestion and accident risk. Road safety research report No 44, Department of Transport, UK

Cambridge systematics (2005) Traffic congestion and reliability. Federal highway administration, Washington, DC

Davidsson P, Henesey L, Ramstedt L, Törnquist J, Wernstedt F (2005) An analysis of agent-approaches to transport logistics. Transp Res Part C 13:255–271

Figliozzi MA (2007) Analysis of the efficiency of urban commercial vehicle tours: data collection, methodology, and policy implications. Transp Res Part B 41:1014–1032

Figliozzi MA (2010) The impacts of congestion on commercial vehicle tour characteristics and costs. Transp Res Part E 46:496–506

Fleischmann B, Gnutzmann S, Sandvoβ E (2004) Dynamic vehicle routing based on on-line traffic information. Transp Sci 38(4):420–433

Fox MS, Barbuceanu M, Teigenm R (2000) Agent-oriented supply chain management. Int J Flex Manuf Syst 12:165–188

Frazzon EM, Makuschewits T, Scholz-Reiter B, Novaes AG (2010) Assessing the integrated scheduling of manufacturing and transportation systems along global supply chains. In: 12th World conference on transport research, paper no 1238, Lisbon, July 11–15

Golden B, Raghavan S, Wasil E (eds) (2008) The vehicle routing problem: latest advances and new challenges. Springer, New York

Goel A (2008) Fleet telematics. Springer, New York
Gumbel EJ (1967) Statistics of extremes. Columbia University Press, New York
Haan L, Ferreira A (2006) Extreme value theory: an introduction. Springer, New York
Isermann R (1997) Supervision, fault-detection and fault-diagnosis methods—an introduction. Control Eng Practice 5(5):639–652
Isermann R (2005) Model-based fault detection and diagnosis—status and applications. Annu Rev Control 29:71–85
Lai TL (2001) Sequential analysis: some classical problems and new challenges. Statistica Sinica 11:303–408
Larsen A (2001) The dynamic routing problem. PhD Dissertation, Technical University of Denmark, Lyngby
Larsen A, Madsen OBG, Solomon MM (2008) Recent developments in dynamic vehicle routing. In: Golden B, Raghavan S, Wasil E (eds) The vehicle routing problem: latest advances and new challenges. Springer, New York, pp 199–218
Mes M, van der Heijden M, van Harten A (2007) Comparison of agent-based scheduling to look-ahead heuristics for real-time transportation problems. Eur J Oper Res 181:59–75
Novaes AG, Burin PJ (2009) A dynamic vehicle routing problem (in Portuguese). In: Proceedings XXIII ANPET—Brazilian association of transport research and teaching, 9–13 Nov, Vitória
Novaes AG, Frazzon EM, Burin PJ (2011) Dynamic vehicle routing in over congested urban areas. In: Kreowski H-J, Scholz-Reiter B, Thoben K-D (eds) Dynamics in logistics. Springer, Berlin Heidelberg, pp 49–58
Psaraftis HN (1995) Dynamic vehicle routing: status and prospects. Ann Oper Res 61:143–164
Ribeiro GM, Lorena L (2005) Dynamic vehicle routing using genetic algorithms (in Portuguese). In: Proceedings of XVI conference on transport research and education —ANPET, Recife
Russel S, Norvig P (2003) Artificial intelligence: a modern approach. Prentice Hall, Englewood Cliffs
Simani S, Fantuzzi C, Patton RJ (2010) Model-based fault diagnosis in dynamic systems using identification techniques. Springer, London
Smith RL (2003) Statistics of extremes, with applications in environment, insurance and finance. Department of Statistics, University of North Carolina, Chapel Hill. (http:www.stat.unc.edu/postscript/rs/semstatrls.ps)
Syslo MM, Deo N, Kowalik JS. (2006) Discrete optimization algorithms with pascal programs. Dover Books on Mathematics, Mineola
Wald A (1947) Sequential analysis. Dover Publications, New York
Weisbrod G, Vary D, Treyz G (2003) Measuring the economic costs of urban traffic congestion to business. Transportation research record no 1839, Transportation Research Board, Washington, DC