Preliminary ideas on a simulation approach to risk assessment of a train timetable.[1]

Arnoldo Frigessi and Magne Aldrin
Norwegian Computing Center, Oslo

Draft: 16.02.1998

Abstract: In this note we analyse the various stochastic components that play a role in a complex railroad network. We describe a simulation based approach to the assessment of the risk of delays and stability of a given timetable. The main methodological keywords are correlations, simulated likelihood, discrete event simulation, space-time analysis, propagation on a network, scenario simulation.

1. Aim

The objective of the project is to create a computational tool which is able to evaluate the risk of delays and stability of a given *ruteplan*.

Stability is the capacity of a *ruteplan* to restore to normal when perturbations, like delays and disruptions, happen. A perturbation is defined as any event that causes trains to behave differently as planned in the time table. Risk is a measure of the cumulated delays of a given *ruteplan* which is perturbed by events causing delays. Risk and stability can refer to the whole network globally or to single elements of it, for instance a region or a type of train, say, only express trains.

The tool will allow to compare different *ruteplan*s in terms of risk and stability.

A by-product of the tool is to help in identifying the problem areas in a given *ruteplan*. It will be possible to suggest and test measures of corrections, of the network and/or of the *ruteplan*, that allow better performances.

The tool will provide a mean to understand the propagation of delays in a network and find relationships between them.

It will allow to test the behaviour of a given *ruteplan* under new assumptions, regarding for instance the equipment, the periodicity of trains or the delay causing events.

[1] We acknowledge helpful discussions with Tor Homleid, VISTA Utredning AS, Oslo.

On the basis of the literature we have seen so far and our experience, we suggest to follow a simulation based approach. Modelling a *ruteplan* on a network analytically or by means of fully descriptive likelihoods is far to complex to deliver reliable results. The structure of a *ruteplan* and a railroad network is so detailed and full of particular exceptions to rules, that only decisions taken on the basis of simulations seem to be trustful. Furthermore simulation is the only way to predict properly the impact on risk and stability of new traffic control policies and equipment.

Our approach can be divided in different tasks:

- understanding and description of the network, equipment and *ruteplan* rules;
- identification of randomness;
- data collection;
- estimation of perturbations and variability;
- simulation procedure;
- computational issues and implementation;
- analysis of output.

In this note we try to follow these steps.

2. Description of a given railroad network.

The analysis must be based on a description of the railroad network. This can be the complete network of NSB or a part of it of particular interest, that is studied independently of the rest (assuming meaningful boundary conditions).

The rail network can be described as a planar graph, where nodes describe stations, crossings, or other important landmarks, and arcs represent sections between such nodes. A section is a collection of physical tracks. Descriptive labels are assigned to each node and arc, so that the relevant technical information is captured. For instance an arc label should contain information on the number of tracks. It may be more convenient to assign one arc and label to each track. We comment later in this section about the concept of section.

There are many aspects that can be taken into account in the description of the network and the level of detail can be tuned to different goals of the analysis.
In principle we think that it is useful to start with a rather complete description of the network, which includes tracks, stations, crossing points, some information regarding the signalling system and maximum speed, etc. Labels of nodes and arcs should contain all such information. It will then be always possible and easy to decide the level of detail by just ignoring or summarising some part of the network information.

An other reason for trying to be rather complete in this descriptive phase, is that this allows us to *learn* about the rail net, and this *learning* will turn to be very important in the modelling tasks. Finally, our impression is that NSB has already a clear picture of its infrastructure and it will be easier and less risky to follow it rather than trying to synthesise.

We will follow very closely the experience of NSB. NSB will show us which are the important features of the network. We are thinking for instance to type and capacity of stations, speed limitations of track segments, type of crossings etc. Properties that should be included are those that

- affect the movements of trains;
- are potentially relevant in creating or solving conflicts between interacting trains.

It is important to distinguish carefully between network characteristics and operational properties. For instance the label of a crossing should include technical constraints regarding the minimal buffer time, but not the actual buffer time that is a quantity that can be decided by NSB and is a characteristic of a specific *ruteplan.* A similar consideration can be made regarding the minimal headway between trains on each track, that has to be declared, while the actual headway decided in a given *ruteplan* does not belong to the label input. Node and arc labels should only incorporate physical, "hard" properties. This distinction is crucial in the simulation phase of the project, where both train movements and traffic control will be simulated given a perturbation of a *ruteplan.* (Recall that we consider every cause of delay or disruption of the normal timetable as a perturbation of it.)

We mentioned the concept of section. We have the impression that NSB has a clear and practical definition, that is related to the actual layout of the network (say, the tracks between two stations on a simple line), but also takes into account requirements coming from the processes of designing and handling a *ruteplan*. What we mean is that probably such segments will not be too long or too complex, because they enter as building blocks into these processes. We think that it will be important to follow the historical experience of NSB and to accept a non rigid logical definition of segment. It could be that a section will be identical to the part of track considered as occupied when a train is using it, and identified by a signal, so that the minimal section is probably a quite small entity.

It is obvious that most of the network description is already done by NSB and that it will be mainly a question of organising the information in a logical and (computationally) convenient way. It is probably practically convenient to start with a subpart of the network that includes Oslo.

3. Description of the rotable equipment.

We need a technical description of the trains, in terms of engines, maximum speeds etc. Again it is important to distinguish such information from actual operation in a given *ruteplan*. It will also be needed to keep a certain generality, in order to avoid overflow of technical information. What is important is that we describe well the main operational restrictions of available trains.

Each train connection in a given *ruteplan* will then be also characterised by the equipment it is composed of.

4. Description of a given *ruteplan* and its randomness.

This is the most complex descriptive and coding part. We first look to single routes of trains.

Routes of trains are modelled as paths in the graph describing the network. The route of a train is described by the sequence of segments it uses, by its departure time from the initial station, by the assumed timing along the route, from one node to the next, by the speed, by the level in train hierarchy (express, local ...) etc. All these properties influence the movements of the train on its route and are also important for traffic control in case of perturbations of the given *ruteplan.* Many, possibly most of these characteristics can be read off from the given *ruteplan* and from the usual time-distance diagrams of the sequence of segments it uses. However it is important to understand and know explicitly which are the principles that lead to a certain timetable, because some of these are likely to be important in the management. and hence in the simulation, of perturbations of the *ruteplan.*

As we have learned, the running time on a timetable of a specific train on a specific route can be considered as an average time, around (or possibly on top of) a theoretical running time, due to several nondeterministic and noncontrollable aspects of such a journey (drivers capacities, number of passengers, meteorological conditions etc.). We will have to address the question of how to cope with this type of randomness. Also here different levels of detail can be modelled.

The FASTA-simulator[2] allows two different operation modes, one called *deterministic mode,* the other *stochastic mode.* We will here explain what is

[2] Mohideen Noordeen, Stability analysis of a cyclic timetable for highly interconnected rail networks, These n. 1435, Ecole Polytechnique Federal de Lausanne, CH, 1996.

meant by these two concepts, but we mention that the term *stochastic* as used by FASTA is restrictive. For us a *stochastic operation mode of the simulation* will be a much more general perturbation of the *ruteplan*, where *events causing perturbations of the regular ruteplan happen non deterministically.* Because we want to study the effects of such events on the running times *globally on the ruteplan,* we have to understand and model the effect of some events explicitly. For instance, we may want to describe carefully the average dwell time at stations and its variability (or distribution) as a function of time of the day (to cope with rush hours etc.) and, say, weather conditions. This will allow us to study, for example, the effect of specific weather conditions during rush hours on the overall delays accumulated in a given *ruteplan*. On the contrary, it is probably difficult to study the impact of the drivers capacities on keeping to scheduled times. These are just examples[3], given in order to make the point that we want to try to describe carefully the randomness inherent in the *ruteplan*, but we are aware that only with the help of experts we will be able to identify the important and measurable sources of variation. (This point has to do also with data collection, and we shall return to this later.)

Back to the restrictive interpretation of *stochasticity* in the FASTA simulation programme. What is meant there is that every train of a certain type has an unexplained random running time, where the distribution of the running time is given in a quite arbitrary fashion. The deterministic operation mode in FASTA is when all trains run at exact *ruteplan* running times.

Both levels of details of the operation mode of a single train should be certainly considered in our project. It will be possible to run the deterministic mode, so that only the events causing delays happen at random. In general the deterministic mode will be less interesting than the stochastic versions, where we think that the distribution of the running times must be estimated from data very carefully, as we will propose later in this note. We are not satisfied with what is implemented in the stochastic mode by FASTA, because it does not allow to identify causes and patterns of propagation of delays.

We now discuss the description of the interaction between trains on their routes.

These can be described as many walks on the network. These walks follow their route, prescribed technical and physical constraints, and try to keep (in both deterministic and stochastic sense) their schedules. Each train is assigned to a segment or arc, and possibly to a specific part of each, if needed. The *ruteplan* relative to a specific segment is well described by the time-distance diagram. Here the interaction between trains is depicted.

[3] As in many parts of this note, examples are probably naïve, possibly also wrong; this is due to our lack of experience in the railroad world. Despite this we hope that the examples are able to illustrate what we mean. We apologise however for their simplicity and, as said, possible incorrectness!

It is useful to know explicitly which are the rules that are used to create these time-distance diagrams, i.e. the principle of the construction of the *ruteplan*. For instance we should know about priorities given to trains of different categories, "who waits for whom", etc.

Together with the technical information relative to the segment and trains (for instance minimum buffer time at crossings), the collection of all diagrams for all segments is the description of the *ruteplan* on the network. Experts will help us to understand all relevant details.

5. Rules of movements in a perturbed *ruteplan.*

We need to know which rules exist and are followed when for some reason the *ruteplan* is perturbed by delays etc. These traffic control rules are important because we have to follow them in the simulations, together with all other technical constraints.

There are probably some actions, taken manually by controllers, like changing the order of trains, that cannot be taken into proper account.

However important strategies used to *recover* after a delay should be modelled, like for instance speeding ups, etc.

Some of these recover rules may be based on short term optimisation of some function counting for delay propagation etc. In fact it may be possible also to incorporate them into the system.

Finally, the rules could be modified from those actually used, in order to test the effect of hypothetical policies. We should allow for flexibility here, thinking to future uses.

6. The simulation of the *ruteplan.*

Once the network, the equipment, the given *ruteplan* and recovery rules are described, a simulator can be run. This means that abstractly trains are started according to timetable, run through the nodes of the network following the rules of routing and interacting with other trains, recovering delays etc.

Because the running time is a random variable, the simulation of a given *ruteplan* must produce random outcomes. We have discussed the fact that we will need do model rather carefully the randomness around the technical running time. The reason is now clear: the simulator must be able to reproduce this variability in a reliable way. This means that distributions of running times and dwell at stations are inputs to the simulation. Hence they must be estimated and well modelled (something not done in FASTA for instance), possibly as functions of other explanatory variables (time of the day, traffic intensity). We shall come back on sampling issues later.

Starting with the earliest train in the morning, the simulation proceeds in time until the end of the day, that we may want to consider as a regeneration point. Technically, a computer system able to perform such a simulation is called *discrete event simulator.* Discreteness refers to the fact that there is a minimal time unit (possibly the second). While the simulation proceeds in time, statistics are collected regarding the performance of the system and these statistics are the primary output of the simulation. Total and partial delays and measures of overall risk and stability are given.

Here a remark is appropriate. If it would be possible to monitor in continuos time the complete network, with all details, for a long time, then these data would be even better than their simulated counterparts. From such complete data one could calculate the delays etc. The point is that for the moment it seems unrealistic to assume the existence of such a controlling and monitoring system. In fact, while some types of data are collected routinely, other that will be needed to model variations will be probably collected on purpose, as we shall discuss later, and will be quite local and partial. It is the task of statistics to recognise inside such partial data general features and patterns.

Building up on these raw simulated data is a further statistical analysis that aims to the understanding of the delay propagation mechanism. These are likely to be more or less standard inferential procedures.

7. Events causing primary delay vs. delay caused by propagation.

We have illustrated the fact that the running time is a random quantity that is estimated by the experts starting from the technical running time. We have explained that we will need to select and identify the mayor causes of such variability. In this section we will discuss in more detail the general issue of causes of delays.[4]

[4] We are not experienced enough to present a complete and clear classification of the causes of delay, and certainly experts will need to help us in this. We will mainly consider some examples in order to clarify the different roles that *logically different* causes of delay play in the simulation of stochasticity.

We understood that there is a crucial distinction between

A. primitive delays, and
B. propagated delays.

A primitive delay is directly caused by an event that is not due to interaction with other trains. For instance in a certain segment a train T breaks and has to slow down, so that it arrives with a certain delay to its next station. There is an event that *originates* the delay. To continue the example, it could be that there are no further train on the segment or needing the segment at the time of the slowing down and that at the arrival station a full recovery can be done. This delay was isolated and did not propagate at all. It is likely however that most delays do propagate. So for instance if a train S has to wait in order to access the segment that is kept occupied for a longer time (due to the event "break down of train T"), then his delay is a propagated one, because the only cause of it, the event that originates it, is simply another delay.

We have seen that to each primitive delay there corresponds an *event originating, causing the delay, and such event is not due to propagation (via interaction) of another delay.*

It is important for us to logically identify the events causing delay that are not simple (!) propagation. A list of such events must be written and data about them collected over time. The time the event happened, the type of event, its location and characteristics should be recorded. From these records we can then estimate distributions of such happenings.

Some of these events are related to equipment, others to the signals, others to stations etc. Accordingly these events will be characterised with certain features. For instance in a station a switch breaks down and hence a platform cannot be used. We would need to know where it happens, possibly which type of switch was involved (but this may be too detailed), what time of the year, how long it took to repair it etc.

We now look to delays caused by propagation. Once an event that causes a primary delay to train T has occurred, the *ruteplan* has to be temporarily modified and all the trains that are affected by the delay of T need to reroute, or wait etc. The rules which are applied automatically or via traffic controllers have been described above and are then applied. Most of these rules are probably deterministic (like wait until a certain segment is released), some may be stochastic (increase speed if the driver feels so).

8. Planning the data collection

In order to estimate the distributions describing the stochastic behaviour of the rail system, data are needed. A lot of data are probably already collected routinely and the first task is to have a clear picture of existing data sources. While ideally the whole network should be monitored for a long enough time, it will be needed to decide that certain segments, trains, stations, etc. are behaving similarly, so that a sample from them only needs to be monitored. We will also assume stationarity in time, with of course periodicities (from one day to the other, or weekly). Stationarity means that we expect that the type of stochasticity of a certain quantity remains unchanged during the time of data collection and in the period the simulation will refer to. This allows to monitor for shorter periods only. The exact assumptions about stationarity in type and time has to be studied and is likely to be based on experience at NSB.

Particularly important are the data regarding the events causing primary delays. These data must be collected and the sample size has to be reasonably large to be able to estimate also tail behaviours, corresponding to extreme cases.

9. Estimation of stochastic behaviours of the network.

We are now facing the most difficult inferential task, the actual estimation of all correlations and distributions given the data collected.

Distributions have been introduced in order to describe the random variability of the rail system. Such distributions have to be modelled, and generally this can be done up to some unknown but crucial parameters, that have to estimated from available data. For instance, one distribution could be related to the running time on a certain track as a function of outer temperature. This function, that links these two variables together, must be decided and generally it is possible to design it up to some unknown parameters, that for instance describe how the speed decreases when the temperature goes below a certain value. The collection of all such distributions, related to the various quantities like for example dwelling times at stations and speed of a specific train on a certain segment, is called stochastic model.

The first important point we want to stress is that there are dependencies among these random quantities that must be modelled correctly. For instance assume that we have two random variables N(A) and N(B), that describe the number of passengers entering a train in station A and in station B, respectively. Then it could be the case that if N(A) is big so is N(B), because of rush hour. Hence it will be important to estimate quantities jointly, because this allows to understand correlations in time, space and type.

This also means that data should be collected in such a way that interactions become measurable. For instance consider the delay due to the event E = "very many passengers entering a train at a specific station". This could be either a primary delay (it is rush hour, call it R) or a propagation delay (a train before has been cancelled, call it C). By keeping statistics of how many times the event E happens together with either R or C, we can understand the exact type of delay and various correlations.

In principle one could write a statistical model for all measured quantities. To do this one would need to model explicitly all correlations. This is a formidable task, and we describe a way to produce estimation without the need of such a general global model.

The idea is called *Simulated Likelihood*[5]and has been developed at NR in the last years. In our context it can be described in the following way. It is easy, or at least feasible, to write a model for certain parts of the global network. For instance we could model the dependency of the dwell time at a station as function of the state of arrival delays at the same station (which causes congestion) as a simple regression problem. There are certainly many parts of the network that we can in fact model well statistically, and the question is then to estimate all unknown parameters in these models. In particular it seems to us reasonable to model all local behaviours well, while it seems more difficult to write an explicit model for quantities related to interaction between trains in movement. If we were to model such global quantities that are linked to each other by means of the propagation dynamics, we would need to write an analytical model for propagation. And the impossibility to do this lead us to simulation.

The collection of well understood and modelled (local) data allows hence to write a partial likelihood. This does not incorporate all information about the system available in the data, because some data have not been used since it is too difficult (or arbitrary) to link them to the other in a *safe* model. If we would estimate parameters on the basis of the partial likelihood alone, first of all we would not use all information which we have. But more important: assume we would now run the simulation using these parameter values and imagine that we would collect a statistics of the delay in Oslo S. during several simulated days. Also assume that in fact data are available about the delay at Oslo S., but that these data were not used in the partial likelihood. If we would compare say the mean delay at Oslo S. as estimated from simulation with the same average delay as estimated from data, we would find a difference. This would not mean that the model of the partial likelihood is wrong, only that we have not used

[5] Tore Schweder, Hans J. Skaug, Mette Langaas, Xeni K. Dimakos, Simulated Likelihood methods for complex double platform line transect surveys, submitted for publication, 1998.

information in full. Hence the question is how to use in full all available information without writing a full complex model. *Simulated Likelihood* is a way to do this. The idea is to estimate the parameters in the partial likelihood (i.e. in the well understood model) in such a way that the simulated global system produces outcomes that are *stochastically coherent* to global additional data.

Hence the estimation procedure is a combination of modelling and likelihood inference with simulation based expected controls. We are very optimistic about the use of this new methodology, that will allow to do coherent and full inference. Simulated likelihood has to be tuned well. But we believe that this is likely to be the correct approach. In fact it is possible to interpret it as an empirical Bayesian approach, where part of the data (those that are difficult to model) are used to design simple priors for the rest.

We mention that we will use careful modelling in the partial likelihoods, including non-parametric and hierarchical Bayesian models. The Bayesian approach allows to incorporate into inference in a mathematical way prior information from data and experts' opinions on the unknown parameters.

10. The simulation of a perturbed *ruteplan.*

Once the stochastic dynamics of the rail system is described and estimated, the simulation will proceed again according to *ruteplan* and rules and every time a random quantity is needed a sample will be drawn from the appropriate distribution. Similarly to section 6, a discrete event simulation is performed. The only important difference is that now samples will be drawn from conditional distributions given all other current variables in the system, i.e. given all possible already incurred perturbations.

The simulation of events causing primary delays is easy. We should think to them as random events happening in time and location according to the estimated (conditional) distributions.

Of course the time needed to perform such a complex simulation can be significant. Different levels of detail and the study of submodels will allow to also control this aspect.

11. Scenario simulation.

What is meant by this is the possibility to apply to a given *ruteplan* simulation specific perturbing events. In the previous section we wrote that events causing primary delays are sampled from a distribution that mimics reality.

In scenario simulation events are positioned deterministically or according to distributions that are not estimated from real data. This could be of interests in order to see the propagation patterns of delays under very special, rare but possible circumstances. For instance, it could be interesting to see the effect on the overall risk of changing the signalling system to a new one that is not going to defect at all, or only very rarely.

Another interesting analysis is to see the effect of simultaneously occurring disruptions, that have not yet been seen together, in order to try to study preventive actions.

It could be interesting to study the effect of other changes in the traffic control. For instance one could be interested in the overall risk if the buffer times are reduced. Or what happens if a new rule is used to recover from delays.

Several interesting analysis can be done with such a simulation tool. However a word of caution is needed. Many parameters of the system have been estimated under the real operation of a specific *ruteplan.* If the scenario applied to the simulation alters significantly the situation, for instance a new type of train is introduced with a very high speed but also high incidence of ruptures, then the unchanged part of the model, that still mirrors the "old" situation, may be wrong. In such heavy extrapolation mode, simulation may bring to misleading results. We will perform sensitivity analysis in order to evaluate how robust certain assumptions are.

12. Simulation analysis: what do we get?

Given a precise definition, the risk of a given *ruteplan* can be measured.

The stability of a given *ruteplan* can be estimated, in terms of recovery time from service perturbations.

Simulation allows to compare the risk of alternative *ruteplans*, or the effect of investment in infrastructure or change in traffic control policy.

The output of the simulation study has to be synthesised and analysed. We will obtain, under the current *ruteplan*, distribution of causes of primary delay, propagation dynamics, statistics regarding the entity of overall and localised delays, etc. These can be specialised so that they regard only a certain category of trains, or a certain period of the day, or a certain equipment, or a certain region, etc.

Probabilities that the delay on certain trains is over a certain threshold can be estimated.

The relative importance of various causes of delay can be identified.

13. Implementation issues.

The programming or modification of a *ruteplan* simulator software is not a simple task and one needs competence that is not statistical and must be acquired.
We think that it would be wise to start from an already available software, like FASTA, or even better RailPlan, or in any case a commercial tool. We are very happy to look into the at NSB already available tools.
We have been looking to FASTA. It has several limitations, but essentially it would be a possible starting point for our activity. We would need to modify it, adding functionalities and more flexible models. It is very likely that other commercial simulators would be similar and we should start by looking to that one already at NSB.

In any case the implementation task relative to simulation can be significant. The computational issues related to inference are also very serious but NR is competent.

It will be important to prepare a project plan that allows for several intermediate objectives. It is probably wise to start with the existing software and level of description of the NSB network and do that parts of statistics that are not based on rewriting the simulator at once. However it seems to us important to have the long term goal in mind from the beginning in order to proceed towards it without too many turns.

14. Some conclusions.

The steps mentioned in the beginning of this note are now clearer:

- understanding and description of the network, equipment and *ruteplan* rules;
- identification of randomness;

- data collection;
- modelling and estimation of partial, local models and integration with global data via simulated likelihood;
- simulation procedure;
- analysis of output.

They represent the backbone of the project. It is possible to first concentrate on a subpart of the network and on a not too fine level of detail, a rougher description of the stochasticity, in order to go through these various step, producing in reasonable time a tool that is useful for risk analysis. It is very important to build up such a large project in order to both obtain important results along the way and aim to a final product that allows to reach the desired aims.