Model predictive control for optimal coordination of ramp metering and variable speed control

A. Hegyi, B. De Schutter, and H. Hellendoorn

If you want to cite this report, please use the following reference instead:

*This report can also be downloaded via https://pub.deschutter.info/abs/01_10.html
Model Predictive Control for Optimal Coordination of Ramp Metering and Variable Speed Control

Andreas Hegyi, Bart De Schutter and Hans Hellendoorn

Control Systems Engineering, Faculty of Information Technology and Systems
Delft University of Technology, P.O. Box 5031, 2600 GA Delft, The Netherlands
Phone: +31-15-278 20 87, Fax: +31-15-278 66 79
email: {A.Hegyi,B.DeSchutter,J.Hellendoorn}@ITS.TUDelft.nl

Abstract
We apply model predictive control to optimally coordinate variable speed limits and ramp metering. The basic idea is that speed limits can increase the (density) range in which ramp metering is useful. For the prediction we use a slightly adapted version of the METANET traffic flow model that takes the variable speed limits into account. The optimal control signals aim at minimizing the total time that vehicles spend in the network. The coordinated control results in a network with less congestion, a higher outflow, and a lower total time spent. In addition, the receding horizon approach of model predictive control results in an adaptive, on-line control strategy that can take changes in the system automatically into account. We illustrate our approach by comparing the cases ‘ramp metering only’ and ‘coordinated ramp metering and speed limits’ for a simple network.

1 Introduction
The steadily increasing number and length of traffic jams on motorways has led to the use of several dynamic traffic management measures all over the world. Usually these measures operate based on local data (occupancy or intensity measurements). However, the last years more and more researchers and practitioners recognize that considering the effect of the measures on the network level has many advantages compared to local control. E.g., solving a local congestion can have as consequence that the vehicles run faster into another downstream congestion because of the improved flow, whereas still the same amount of vehicles have to pass the bottleneck (with given capacity). So, the average travel time in the network level will still be the same. Another reason is that in a dense network the effect of a local control measure can have effects on more distant parts of the network: an improved/delayed flow could cause/prevent congestion somewhere else in the network. Furthermore, if dynamic origin-destination (OD) data is available, control on the network level can take advantage of the predicted flows in the network. Local controllers are not able to use OD information because the flow arriving at the local controller depends on the actions of other controllers in the network, which are unknown. E.g., during peak hours the density on the mainstream (motorway) can be so high that the queue on an on-ramp spills back to the surface streets of the city, whereas (pro-active, coordinated) metering of upstream on-ramps could reduce the density of the mainstream flow and prevent spilling back of the on-ramp queue. Another source of degradation of network performance is that congestion might block traffic directions that have nothing to do with the congestion, such as blocked motorways because of congested off-ramps, or blocking via motorway intersections.

To solve these problems a control strategy on the network level is needed. We mean by ‘network level’ the network-wide coordination of control measures, i.e., the measures are operated based on global data. Besides using global data a prediction of the network evolution is also needed to achieve optimal network control, since the effect of a control measure on more distant locations will only be visible after some time. To predict the effects of a control measure several techniques can be used, such as case-based reasoning [4], rule-based systems [2], or model-based prediction. In this paper we use the METANET traffic simulation model [5, 10] for the predictions, and apply a model predictive control framework [1,3] to find the optimal combination of control measures (control inputs). This framework requires an objective function that expresses the performance of the traffic network (as a function of a given control input). The goal is to minimize the total time spent (TTS) by the vehicles in the network. Papageorgiou [8] showed that, under the condition that the network inflow is known or can be predicted accurately, minimizing
TTS is equivalent to maximizing the outflow of the network weighted by the factor \( K - k \), where \( k \) is the time index and \( K \) the last time instant of the period in which the network evolution is considered. That means that a controller with this objective function (minimize TTS) will tend to maximize the outflow as soon as possible. We will add another term to the objective function to discourage abrupt changes in the control signal.

Ramp metering can be used for two different purposes. First, when drivers try to bypass congestion on a local road (rat running), ramp metering can increase travel times and discourage the use of the bypass [7]. Second, when traffic is dense, ramp metering can increase flow or prevent a traffic breakdown by adjusting the metering rate such that the density remains below the critical value. Several field and simulation studies have shown the effectiveness of ramp metering in the latter sense [11, 12, 14, 15]. Generally a 0–5 % increase of the flow and a 0–10 % increase of the speed is achieved. In a simulation study Kotsialos et al. [6] reported a decrease of 20–30 % of the TTS by optimal coordinated ramp metering on the Amsterdam ring road. In this paper we focus on the second objective of ramp metering, i.e., improving the traffic flow.

Variable speed limits are much less known and studied as a tool to reduce congestion. More known effects of speed limits are homogenization and increase of safety [13, 16]. In these studies Smulders applied the same speed limit to each section of a longer motorway stretch, by which the variance of the speed is shown to decrease. Another application of speed limits is queue warning. In the Netherlands such a system is active: it sets appropriate speed limits on upstream sections if an unusual situation (incident, low speed) is detected. To the authors' best knowledge there is no publication about applying dynamic speed control to reduce or prevent congestion. However, the Dutch Ministry of Transport, Public Works and Water Management is preparing an experiment in the DYVERS project where the reduction of congestion by dynamic speed limits is studied.

In this paper we use a combined approach in which ramp metering and dynamic speed limits are coordinated to increase the range in which ramp metering is useful.

![Figure 1](image-url) Figure 1: If traffic on the main road is in state 1, it is nearly unstable and even a small flow from the on-ramp may cause a breakdown. Speed limits change the state from 1 to 2, and change the shape of the fundamental diagram from the solid gray line to the dashed black line. This increases the stable region and creates space for the on-ramp traffic. The additional vehicles from the on-ramp change the state from 2 into the direction of 3.

2 Problem description

It is clear that ramp metering is only useful when traffic is not too light (otherwise ramp metering is not needed) and not too dense (otherwise breakdown will happen anyway). This region is on the stable (no breakdown) side of the fundamental diagram, and close to the top (see Figure 1), because that is where a breakdown can happen.

The main idea of this paper is that the combination of ramp metering with variable speed limits increases the (density) range where ramp metering is useful. This prevents or postpones a traffic breakdown when traffic is getting dense. It is important to note that the congestion after a breakdown has usually a flow that is 5–10 % lower than the available capacity. Papageorgiou [9] showed that a decrease of outflow of 5 % can result in an increase of total time in the network of 20 %. The reason for this big difference is that the queue integrates the difference between the inflow and the outflow. But the outflow is lower when there is congestion (capacity-drop phenomenon), so the queue grows faster, and consequently congestion will last longer, and the outflow will be low for a longer time. This is why one should try to prevent or postpone a breakdown as much as possible.

In Figure 2 speed limits are imposed on the mainstream traffic while the on-ramp is metered. The speed limits change the shape of the fundamental diagram (see Figure 1). The reason why speed lim-
its increase the range where ramp metering is useful is the following. Suppose that the flow on the main road is close to capacity and that the density is close to the critical density (state 1 in Figure 1). In this state even a small flow from the on-ramp can cause a breakdown. Now we apply speed limits of 70 km/h to the two sections of the main road. The density in these sections will remain approximately the same, and the drivers will experience relatively large headway distances (approximately the inverse of the density), because their headway distance was chosen to match a higher speed. So, the state in Figure 1 is changed from 1 to 2. This ‘increased’ headway distance can be used to accommodate more traffic from the on-ramp. The traffic flow entering from the on-ramp will shift the state from 2 into the direction of 3, and the flow on the main road will remain stable for a larger range of on-ramp flows.

A drawback is that the flow will also decrease by the same factor as the speed. But if the control is optimized properly, this flow drop will be always less than or equal to the flow drop of a breakdown, since otherwise breakdown would be the optimal situation. Another point of criticism could be that the approach would keep the controlled network congestion free, but at the cost of creating congestion at the entrances of the network. This is only partially true, because the controller will indeed sometimes delay the traffic to prevent a breakdown in the network, but afterwards the flow will be higher than if the breakdown would have occurred. So the inflow of the network will be decreased by the speed limits only for a short period of time. Unfortunately this still can cause congestion on upstream sections, but if this would happen two things can be done. First, the region of the network that is used to evaluate the performance can be chosen larger than the region that is controlled, so that effects outside the controlled region are also taken into account. Second, the network that is considered (evaluated and controlled) can be chosen larger, because the traffic is apparently so dense that the effects of the control should reach beyond the bounds of the actual network.

3 Approach

3.1 Model Predictive Control

To solve the problem of coordination of speed limits and ramp metering we apply a model predictive control scheme in a rolling horizon framework [1,3]. Every time step $k$ the optimal control inputs are found (by optimization) for the state of the network at time $k$ and the expected or predicted demand over a given prediction horizon $N_p$. In order to limit the number of variables (control inputs) in the optimization procedure and to improve the stability of the system, we also introduce a control horizon $N_c (< N_p)$: after the control horizon has been passed the control signal is taken to be constant. The optimization procedure includes a prediction of the network evolution as a function of the current state and a given control input (this prediction is based on a slightly modified version of the METANET model — see below). In that way we can optimize a performance measure that in our case is basically equal to the TTS. Next, the controller executes the first step of the optimal control input, and the whole procedure is repeated due to the rolling horizon approach. This rolling horizon approach results in an on-line adaptive control scheme that allows us to take changes in the system or the system parameters into account by regularly updating the model of the system or the predicted demands as new measurements become available.

Important parameters in this scheme are the lengths of the control and prediction horizons. Too long control and prediction horizons can result in intractable optimization problems (recall that the optimization has to be performed on-line, so the size of the problem, i.e., the number of variables and the length of the prediction, should not be too large). On the other hand, the prediction horizon $N_p$ has to be long enough to represent the important process dynamics, and the control horizon $N_c$ has to be long enough to be able to achieve a reasonable performance. In conventional model predictive control [1,3] heuristic tuning rules have been developed to select appropriate values for $N_p$ and $N_c$. Based on heuristic reasoning we can expect that the optimal value of the prediction horizon $N_p$ is about the typical travel time in the network. If the prediction time is taken shorter, then the effect of the vehicles that are influenced by the current control measure and — as a consequence — have an effect on the network performance before they exit the network, will not be taken into account. If the prediction horizon is longer than the typical travel time, the affected vehicles will exit the network before the end of the
prediction horizon and may have no effect on the network performance. On the other hand, a control action may affect the network state (by improved flows, etc.) even when the actually affected vehicles have already exited the network. For the control horizon \(N_c\) we will select a value that represents a trade-off between the computational effort and the performance. In our simulations we will also investigate the effect of changing the length of the control and prediction horizons. This should result in heuristic rules of thumb for the selection of \(N_p\) and \(N_c\) for model predictive traffic control.

### 3.2 Model

For predicting the future effects of a given control signal, we use the METANET model [5]. Only those parts of the model that are used in our simulations are described here. The METANET model represents a network as a directed graph whereby the links (index \(m\)) represent motorway stretches. Where major changes occur in the characteristics of the link or in the road geometry (on/off-ramp), a node is placed. Each link is divided into segments (index \(i\)) of length \(L_m\). Each segment is characterized by the traffic density \(\rho_{m,i}(k)\) (veh/lane/km), the mean speed \(v_{m,i}(k)\) (km/h), and the traffic volume or flow \(q_{m,i}(k)\) (veh/h), where \(k\) indicates the time instant \(t = kT\), \(k = 0, \ldots, K\), and \(T\) is the discrete time step (typically \(T = 10\) s). The following equations describe the evolution of the network state over time. The outflow of each segment is equal to the density multiplied by the mean speed and the number of lanes on that segment:

\[
q_{m,i}(k) = \rho_{m,i}(k) \cdot v_{m,i}(k) \cdot L_m, \tag{1}
\]

where \(L_m\) is the number of lanes on link \(m\). The density of segment \(i\) of link \(m\) at time step \(k + 1\) equals the density at time step \(k\) plus the inflow from the upstream segment, minus the outflow of the segment itself (law of conservation of vehicles):

\[
\rho_{m,i}(k+1) = \rho_{m,i}(k) + \frac{T}{L_m \cdot L_m} (q_{m,i-1}(k) - q_{m,i}(k)). \tag{2}
\]

The mean speed at time step \(k + 1\) equals the mean speed at time \(k\) plus a relaxation term that expresses that the drivers try to achieve their desired speed \(V(\rho)\), a convection term that expresses the speed increase (or decrease) caused by the inflow of vehicles, and an anticipation term that expresses the speed decrease (increase) as drivers experience a density increase (decrease) downstream:

\[
v_{m,i}(k+1) = v_{m,i}(k) + \frac{T}{\tau} \left( V(\rho_{m,i}(k)) - v_{m,i}(k) \right) + \frac{T}{L_m} \left( v_{m,i-1}(k) - v_{m,i}(k) - \frac{\nu T}{\tau L_m} \left( \rho_{m,i+1}(k) - \rho_{m,i}(k) \right) \right), \tag{3}
\]

where \(\tau, \nu\) and \(\kappa\) are model parameters.

The desired speed \(V\) is a function of the density \(\rho_{m,i}(k)\):

\[
V(\rho_{m,i}(k)) = \nu_{\text{free},m} \exp \left[ -\frac{1}{a_m} \left( \rho_{m,i}(k) \right)^{a_m} \right], \tag{4}
\]

where the free speed \(\nu_{\text{free},m}\), the critical density \(\rho_{\text{crit},m}\), and the parameter \(a_m\), are model parameters. The physical interpretation of \(\nu_{\text{free},m}\) is the average speed that drivers assume if traffic is flowing freely, and \(\rho_{\text{crit},m}\) is the density at which the traffic flow becomes unstable (cf. Figure 1).

Besides the motorway links, a special link is used to model origin (on-ramp) links, which receive traffic demand and forward it to the motorway. A simple queue model is used to describe the dynamics of the origin link. The length of the queue \(o(k + 1)\) in the origin link \(o\) equals the previous queue length plus the demand \(d_o(k)\), minus the outflow \(q_o(k)\):

\[
o(k+1) = o(k) + T(d_o(k) - q_o(k)). \tag{5}
\]

The outflow of the origin depends on the traffic conditions on the mainstream and the ramp metering rate \(r_o(k)\), where \(r_o(k) \in [0, 1]\). More specifically, \(q_o(k)\) is the minimum of two quantities: the demand in time period \(k\) plus the queue on the on-ramp, and the maximal flow that can enter the motorway because of the mainstream conditions, modified by the metering rate:

\[
q_o(k) = \min \left[ d_o(k) + \frac{w_o(k)}{T}, Q_o \min \left( r_o(k), \frac{\rho_{\text{max}} - \rho_{\text{crit},m}}{\rho_{\text{max}} - \rho_{\text{crit},m}} \right) \right], \tag{6}
\]

where \(Q_o\) is the on-ramp capacity (veh/h) under free-flow conditions and \(\rho_{\text{max}}\) (veh/lane/km) is the maximum density, and \(\rho\) the index of the link to which the on-ramp is connected. In order to account for the speed drop caused by merging phenomena, if there is an on-ramp then the term \(-\delta T q_o(k) v_{m,1}(k) / L_m A_m (\rho_{m,1}(k) + \kappa)\) is added to
(3), where $\delta$ is a parameter, $\mu$ is the index of the merging link, and $m$ is the index of the leaving link.

Since in order to evaluate the evolution equations for a segment we need an upstream speed and flow, and a downstream density, the nodes (that connect the links) in the network provide the entering and leaving links (the last and first segments) with the appropriate values. In our case the speed of the last segment of the entering link is simply passed to first segment of the leaving link

$$v_{m,0}(k) = v_{\mu,N_{\mu}}(k),$$

where $m$ is the leaving link, $\mu$ the entering link, and $N_{\mu}$ the index of the last segment of link $\mu$.

Furthermore, the sum of the flows of the entering links equals the inflow of the leaving link:

$$q_{m,0}(k) = q_{\mu,N_{\mu}}(k) + q_o(k),$$

where $q_o(k)$ is the flow from the on-ramp (if there is one) connected to the node, and $N_{\mu}$ is the index of the last segment of link $\mu$. The downstream density of the last segment $N_{\mu}+1$ of link $\mu$ is the density of the first segment of the leaving link $m$:

$$\rho_{\mu,N_{\mu}+1}(k) = \rho_{m,1}(k).$$

Since the original METANET model does not describe the effect of speed limits we have slightly modified the desired speed equation (4) to incorporate speed limits: we assume that the desired speed is always less or equal to the speed limit $v_{\text{control}}(k)$ displayed on the variable message sign (VMS):

$$V(\rho_{m,i}(k)) = \min \left( v_{\text{control}}(k), \right.$$

$$v_{\text{free},m} \exp \left[ -\frac{1}{a_m} \left( \frac{\rho_{m,i}(k)}{\rho_{\text{crit},m}} \right)^{a_m} \right] \left. \right).$$

In addition, for $O_1$ we use a modified version of (6) with another flow constraint, because when a speed limit is active, the inflow can never be higher than the highest flow of the fundamental diagram with the given speed. So

$$\dot{q}_o(k) = \min \left[ d_o(k) + \frac{w_o(k)}{T}, \Lambda_m v_{\text{control}}(k) \right.$$

$$\rho_{\text{crit},m} \left( -a_m \ln \left( \frac{v_{\text{control}}(k)}{v_{\text{free},m}} \right) \right)^{\frac{1}{a_m}} \left. \right].$$

3.3 Objective function

The model predictive control algorithm finds the control signal that minimizes the following objective function:

$$J = \sum_k \left\{ \sum_{m,i} \rho_{m,i}(k) L_m \Lambda_m + \sum_o w_o(k) + \right.$$

$$a_{\text{ramp}} \left( r_o(k) - r_o(k-1) \right)^2 +$$

$$a_{\text{speed}} \left( v_{\text{control}}(k) - v_{\text{control}}(k-1) \right)^2 \right\}. $$

This objective function contains two terms for the TTS (one term for the mainstream flow and one term for the on-ramp queue), and two terms that penalize abrupt variations in the ramp metering and speed limit control signals respectively. These terms are weighted by the nonnegative weight parameters $a_{\text{ramp}}$ and $a_{\text{speed}}$.

In order to illustrate the control framework presented above we will now apply it to a simple traffic network.

4 Simulations

4.1 Network and scenarios

The network for the experiments (Figure 3) was chosen as simple as possible. It basically consists of a main stream with speed limits, and a metered on-ramp. The choice for the second speed limit was made to have more control over the state (speed, density) in the segment that is just before the on-ramp. The network considered consists of two origins, two motorway links, and one destination. $O_1$ is the main origin and has two lanes with a capacity of 2000 veh/h each. The motorway link $L_1$ follows with two lanes, and is 2 km long consisting of two segments of 1 km each. Both segments are equipped with a VMS where speed limits can be set. At the end of $L_1$ a single-lane on-ramp ($O_2$) with a capacity of 2000 veh/h is attached. Link $L_2$ follows with two lanes and a length of 1 km, and ends in destination $D_1$ with unrestricted outflow. We assumed that the queue at $O_2$ may not exceed 100 vehicles,
in order to prevent blocking back to a surface street intersection.

We have used the network parameters as found in [5]: $T = 10$ s, $\tau = 18$ s, $\kappa = 40$ veh/lane/km, $v = 60$ km$^2$/h, $\rho_{\text{max}} = 180$ veh/lane/km, $\delta = 0.0122$, and $\rho_{\text{crit}} = 33.5$ veh/lane/km. To examine the effect of the combination of variable speed limits and ramp metering a typical demand scenario was considered (see Figure 4): The mainstream demand has a constant, relatively high level, while the demand on the on-ramp increases to near capacity, remains constant for a half an hour, and decreases finally to a constant low value. This scenario is chosen such that congestion sets in at a certain moment and resolves some time later. The important quantities are the TTS (related to the outflow), and the time instant where congestion sets in. For this scenario these quantities will be compared for the ‘coordinated speed limits and ramp metering’ (CSLRM), and the ‘ramp metering only’ (RM) case.

### 4.2 Results

The optimal prediction horizon was found to be approximately $N_p = 42$ (7 minutes), which is in the order of the typical travel time through the network ($4$ km / $40$ km/h). Shorter prediction horizons did not take the whole response of the system into account and resulted in insufficient control actions. Longer prediction horizons tended to take the future demand too much into account, which degraded the performance. For the control signals we have assumed that they can change only every minute, which is more realistic than every 10 seconds. A control horizon $N_c = 3$ (3 minutes) was sufficient for the ‘ramp metering only’ case, for the coordinated speed limit and ramp metering case a control horizon of $N_c = 5$ was necessary. Longer control horizons tended to give the control signal optimization too much freedom, which resulted in more variance in the signals.

The results of the two cases are shown in Figures 5 and 6. The time that both queues at the on-ramps are resolved is somewhat less than $3$ hours in the RM case, and $2$ hours in the CSLRM case. The TTS in the RM case was $815$ veh.hours and in the CSLRM case $737$ veh.hours, which is an improvement of nearly $10\%$.

### 5 Further research

Most of the topics for further research are related to modeling. In (10) and (11) we have modeled the effect of a speed limit as a change in the desired speed of the drivers. We have implicitly assumed that this change occurs for all the drivers, which may not reflect reality, because only those drivers that have entered the segment after the speed limit has changed, will change their desired speed. Furthermore, we have assumed that the drivers fully comply with the displayed speed limits. A stochastic distribution may be a better way of modeling of compliance.

Whichever model is chosen for the speed limits, the model has to be calibrated and validated with real data. It is reasonable to expect that calibration with different (real) data will result in different model parameters. In future, the effectiveness of model predictive control for optimal coordination of speed limits and ramp metering with different model parameters will be studied. It is expected that this model predictive coordination of control measures will improve performance for every set of model parameters that result in a model that can reproduce the capacity-drop phenomenon.

In our simulations we have assumed that the real world coincides with the model used by the controller to make predictions, and that the disturbances (demands) are exactly known. In practice, the model of the controller is always different from the real traffic system, and the disturbances are only partially known. Model predictive control is known to perform well when this kind of differences occurs.

Another interesting and relevant topic is the study of larger networks, such as ring roads around cities or areas where several cities are connected by motorways and OD demands go typically through the whole network. Also considering other control measures than speed limits and ramp metering (such as
Figure 5: The simulation results of the 'ramp metering only' case. When the density approaches the critical density (33.5 veh/km) the ramp metering gradually switches on, and keeps the flow high (around 4200 veh/h). After the queue length on the on-ramp has reached the maximum queue length (100 veh), the ramp metering is forced to let the complete demand enter the highway. This causes a congestion, and in consequence a reduced outflow. The queue at O₁ is only resolved after 3 hours (not shown here).

Figure 6: The results of the coordinated speed limits and ramp metering case. The beginning of the ramp metering signal is similar to the ramp metering only case, but when the density becomes too high (which endangers the high outflow) the speed limits switch on and reduce the inflow from the mainstream. This, of course, causes a longer queue on the mainstream, but the outflow is kept higher and both queues (O₁ and O₂) are resolved within less than two hours.
peak-lanes, route information, reversible lanes, etc.) is a topic for future research.

In the objective function in this paper the same weight was used for the time that vehicles spend on the on-ramp as the time that they spend on the motorways. By putting a higher weight on one of the terms, a preference can be expressed for long-distance (or short-distance) traffic. E.g., weighting the time spent on the motorway more, will result in a shorter average travel time on the motorways, but a longer waiting time on the on-ramps. This will encourage the short-distance drivers to choose another route. Future research will be directed at the study of the effect of different weights on a network where short-distance and long-distance traffic can be distinguished.

6 Conclusions

In this paper we have applied model predictive control to optimally coordinate variable speed limits and ramp metering. The combination of speed limits with ramp metering increased the (density) range in which ramp metering is useful. This idea is illustrated by a simple example network, where the cases ‘ramp metering only’ and ‘coordinated ramp metering and speed limits’ are compared for several demand scenarios. For both cases under all scenarios the control signal was optimized such that the total time spent in the network is minimal. It was found that the coordinated case results in congestion-free network that has a higher outflow and a lower TTS.

References