periodica polytechnica

www.pp.bme.hu

Transportation Engineering

2000 28/1-2

PB2002-104402

REPRODUCED BY: NTIS
U.S. Department of Commerce
National Technical Information Service
Springfield, Virginia 22161

Budapest University of Technology and Economics

periodica polytechnica

A contribution to international technical sciences, published by the Budapest University of Technology and Economics in six separate series, covering the following sciences:

Chemical Engineering
Civil Engineering
Electrical Engineering and Informatics
Mechanical Engineering
Social and Management Sciences
Transportation Engineering

UNIVERSITY LEADERS:

Á. DETREKŐI, Rector Magnificus
GY. HORVAI, Vice Rector for Research and International Relations
F. VÖRÖS, Vice Rector for Education
B. PETRÓ, Dean of the Faculty of Architecture
M. KUBINYI, Dean of the Faculty of Chemical Engineering
GY. FARKAS, Dean of the Faculty of Civil Engineering
L. PAP, Dean of the Faculty of Electrical Engineering and Informatics
K. MOLNÁR, Dean of the Faculty of Mechanical Engineering
T. KESZTHELYI, Dean of the Faculty of Natural Sciences
J. KÖVESI, Dean of the Faculty of Economic and Social Sciences
É. KÖVES-GILICZE, Dean of the Faculty of Transportation Engineering

EDITORIAL BOARD OF PERIODICA POLYTECHNICA:

FERENC WETTL, General editor
MÁRIA TARJÁN-FÁBRY, Technical editor

Transportation Engineering

Published twice a year

SCIENTIFIC ADVISORY BOARD:

Chairman: B. KULCSÁR, Vice Dean for Scientific Research
Members: É. KÖVES-GILICZE, J. BOKOR, J. MÁRIALIGETI, I. ZOBORY, J. TAKÁCS

EXECUTIVE EDITORIAL BOARD:

Head: K. KURUTZ
Members: L. KATKÓ, A. SZABÓ, T. PÉTER

Contributions in Mechanical Engineering should be sent to the *address of the journal:*

Mail address: Budapest University of Technology and Economics
PERIODICA POLYTECHNICA, Transportation Engineering
1521 Budapest, HUNGARY
Fax: 36 1 463-2270;
E-mail: perpol@eik.bme.hu

For *subscriptions email to* order@kkt.bme.hu, for *exchange* copies email to exchange@kkt.bme.hu, or in both cases mail to the Acquisition Department of the Central Library of the Budapest University of Technology and Economics (H-1111 Budapest, Budafoki út 4.; Fax: 36 1 463-2440).
Published by the Budapest University of Technology and Economics, Budapest, Hungary, with the financial help of the foundation "Ipar a korszerű mérnökképzésért" (Industry for Modern Education in Engineering).

HU ISSN: 0303-7800 (online version: HU ISSN 1587-3811)

periodica polytechnica
www.pp.bme.hu

Transportation Engineering

2000 28/1-2

Budapest University of Technology and Economics

PRINTED IN HUNGARY
LIGATURA LTD – NASZÁLY PRINT LTD, VÁC

PERIODICA POLYTECHNICA SER. TRANSP. ENG. VOL. 28, NO. 1–2, PP. 3–16 (2000)

MIXED $\mathcal{H}_2/\mathcal{H}_\infty$ CONTROL DESIGN FOR ACTIVE SUSPENSION STRUCTURES

Péter GÁSPÁR*, István SZÁSZI†, József BOKOR*

*Computer and Automation Research Institute, Hungarian Academy of Sciences
Kende u. 13-17, H-1111 Budapest, Hungary
Fax : +36 14667503 and e-mail: gaspar@sztaki.hu

†Department of Control and Transport Automation, Budapest
University of Technology and Economics
H–1111 Budapest, Műegyetem rkp. 3, Hungary
Fax: +36 1 463-3087, e-mail: szaszi@kaut.kka.bme.hu

Received: Nov. 1 , 2000

Abstract

This paper presents the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ synthesis for active suspension design. The compensator is designed in such a way that it minimizes a given $\mathcal{H}_2$ performance function, which keeps the maximum supported $\mathcal{H}_\infty$ perturbation below appropriate levels. In this design problem, the aim is to find a set of values for the design variables that yield an optimum value of the objective (or cost) function and which fit a number of constraints. In order to solve this problem a design procedure based on a trade-off curve is presented.

Keywords: mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control, robust control, optimal control, flexible structures, uncertain linear systems.

1. Introduction

The aim of the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control synthesis is to treat the standard $\mathcal{H}_2$ and $\mathcal{H}_\infty$ optimal control problems as separate problems but in a unified state-space framework. This method provides a compensator that combines the $\mathcal{H}_2$ quadratic performance criterion for disturbance rejection with the $\mathcal{H}_\infty$ performance criterion for maximum robustness against destabilizing uncertainties. It means, the controller which minimizes the $\mathcal{H}_2$ performance index is selected from the suitable $\mathcal{H}_2$ controllers [1, 3, 13, 15]. The solution of the optimization problem leads to three Riccati equations, which are mutually interconnected, and therefore the solution can be reached in an iterative way. There are effective and powerful algorithms to solve this problem, e.g. the homotopy technique, or a quasi-Newton technique [12, 11]. Another solution is based on the matrix inequality method, i.e. the interior point algorithm [10].

The purpose of the active suspension is to eliminate the harmful vibration caused by road irregularities and on-board excitation sources. However, the model contains uncertainties, which are caused by the actuator error, parameters varying around their nominal value, neglected non-linear effects, and uncertain components.

Several methods have been proposed for solving the active suspension design problem. The $\mathcal{H}_2$ control is well suited to the design of nominal performance in terms of disturbance rejection, however, performance cannot be guaranteed in the presence of uncertainties [6, 8]. The $\mathcal{H}_\infty$ control guarantees robust stability and nominal performance in the presence of uncertainties, however, it often results in a conservative controller [14, 7]. In this paper we concentrate on the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ method, in which the conservatism of the $\mathcal{H}_\infty$ method is reduced.

The paper is structured as follows. Section 2 presents the suspension structures which are used in the design process. Section 3 shows the principle of the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control design, and it discusses the solution of the control synthesis. In Section 4 the method is illustrated through simulation results. Finally, Section 5 presents concluding remarks.

2. Models Used in the Suspension Design

Suspension performance criteria include wheel-load variation, static and dynamic attitude control, working space, discomfort, and steering behavior. Using a half-car model, wheel-load variation, suspension working space and body mass acceleration can be calculated. Since the displacement of the body mass can be linked with discomfort, the criteria above can be considered applying a half-car model [2, 4, 5, 9].

The well-known rigid half-car vehicle model, which is shown in *Fig. 1*, is widely used for active suspension design. The model comprises three parts: the sprung mass and two unsprung masses. Let the sprung and unsprung masses be denoted by m_s and m_a, respectively. Both suspensions consist of a linear spring, a damper and an actuator to generate a pushing force between the body and axle. The front and rear suspension stiffness, the front and rear tire stiffness are denoted by k_{sf}, k_{sr} and k_{tf}, k_{tr}, respectively. The front and rear suspension damping are denoted by b_{sf}, b_{sr}. The half-car model has four degrees-of-freedom: The sprung mass is assumed to be a rigid body and has freedoms of motion in the vertical and pitch direction. Each of the unsprung masses has freedom of motion in the vertical direction. Let the front and rear displacement of the sprung and the unsprung mass be denoted by x_{1f}, x_{1r} and x_{2f}, x_{2r}. In the half-car model, the disturbances, w_f, w_r are caused by road irregularities. The input signals, u_f, u_r are generated by the actuators.

The state space representation (SSR) of the half-car model can be formalized as follows:

$$\dot{x} = Ax + B_1 w + B_2 u, \tag{1}$$

where the state, disturbance and input force vectors and the system matrices are the following:

$$x = \begin{bmatrix} x_1 & \theta & x_{2f} & x_{2r} & \dot{x}_1 & \dot{\theta} & \dot{x}_{2f} & \dot{x}_{2r} \end{bmatrix}^T ;\ w = \begin{bmatrix} w_f & w_r \end{bmatrix}^T ; u = \begin{bmatrix} u_f & u_r \end{bmatrix}^T ,$$

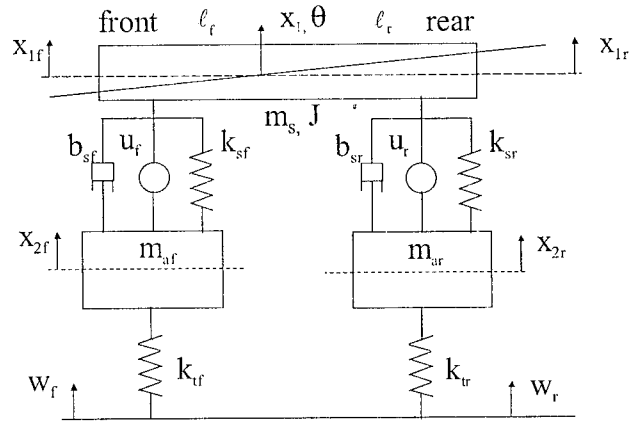


Fig. 1. Half-car model with rigid body structure

$$A = \begin{bmatrix} 0 & I \\ -M^{-1}K & -M^{-1}B \end{bmatrix}; \; B_1 = \begin{bmatrix} 0 \\ -M^{-1}K_r \end{bmatrix}; \; B_2 = \begin{bmatrix} 0 \\ -M^{-1}G_a \end{bmatrix}.$$

Here the sprung mass, the unsprung mass, the suspension stiffness, the tire stiffness, and the suspension damping matrices are as follows:

$$M = \begin{bmatrix} M_s & 0 \\ 0 & M_u \end{bmatrix}; \quad B = \begin{bmatrix} GB_sG^T & -GB_s \\ -B_sG^T & B_s \end{bmatrix};$$

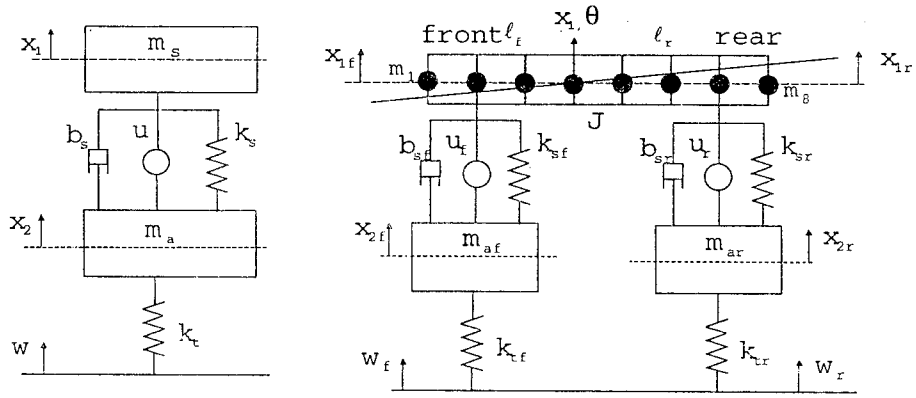
$$K = \begin{bmatrix} GK_sG^T & -GK_s \\ -K_sG^T & K_s + K_t \end{bmatrix}; \quad K_r = \begin{bmatrix} 0 \\ K_t \end{bmatrix}; \quad G_a = \begin{bmatrix} -G \\ I \end{bmatrix},$$

where

$$M_s = \begin{bmatrix} m_s & 0 \\ 0 & J \end{bmatrix}; \quad B_s = \begin{bmatrix} b_{sf} & 0 \\ 0 & b_{sr} \end{bmatrix}; \quad K_s = \begin{bmatrix} k_{sf} & 0 \\ 0 & k_{sr} \end{bmatrix}$$

$$M_u = \begin{bmatrix} m_{af} & 0 \\ 0 & m_{ar} \end{bmatrix}; \quad K_t = \begin{bmatrix} k_{tf} & 0 \\ 0 & k_{tr} \end{bmatrix}; \quad G = \begin{bmatrix} 1 & 1 \\ l_f & -l_r \end{bmatrix}.$$

If other suspension models are applied, their SSR can be formalized in a similar way. A simpler structure is the so-called quarter-car model, which is shown on the left hand side of *Fig. 2*, in which the pitch angle cannot be taken into consideration. Another way to model the suspension structure is to use a flexible model, in which the real situations can be analyzed. The flexible model is shown on the right hand side of *Fig. 2*.



(a) Quarter-car model with rigid body structure

(b) Half-car model with flexible body structure

Fig. 2. Model structures for active suspension design

3. Control design based on the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ method

Consider the linear plant G with input u, disturbance r, performance outputs z_∞ and z_2, feedback output y. The input is generated by output feedback, using the controller K. The signal z_∞ is the performance associated with the $\mathcal{H}_\infty$ constraint, the signal z_2 is the performance associated with the $\mathcal{H}_2$ criterion. The state space representation (SSR) of the controlled system can be written as follows:

$$\begin{aligned} \dot{x} &= Ax + B_1 w + B_2 u \\ z_\infty &= C_1 x + D_{12} u \\ z_2 &= C_2 x + D_{22} u \\ y &= C_3 x + D_{31} w, \end{aligned} \tag{2}$$

where (A, B_2) is assumed to be stabilizable and (A, C_3) is assumed to be detectable. These conditions ensure the existence of stabilizing controllers, and the existence of a K that stabilizes the $\mathcal{H}_2$ problem has been shown to be necessary and sufficient for K stabilizing the $\mathcal{H}_\infty$ problem. The desired compensator K can be determined from an optimization problem. The illustration of the controlled system is shown in *Fig. 3*.

$$\begin{aligned} \dot{x}_c &= A_c x_c + B_c y \\ u &= C_c x_c. \end{aligned} \tag{3}$$

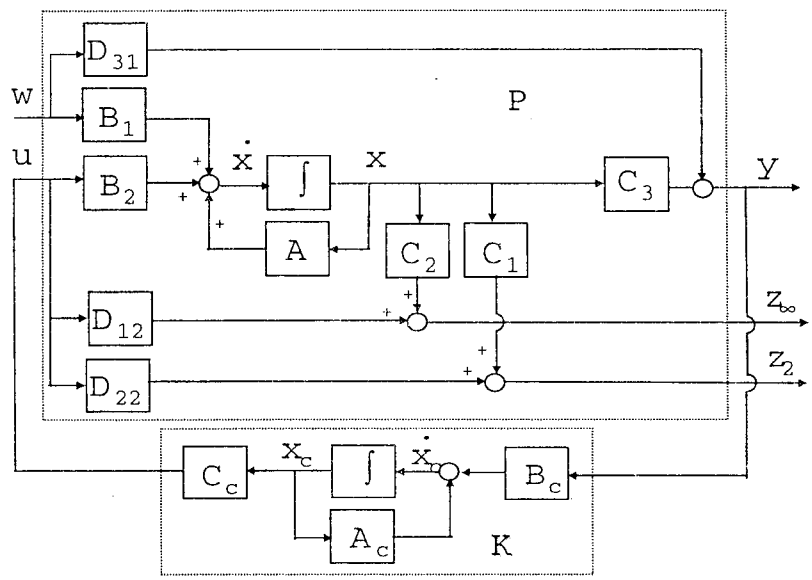


Fig. 3. The closed-loop system for the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control design

The objective of mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control is to minimize the 2-norm of the closed-loop transfer function $T_{z_2 w}$, while constraining the inf-norm of the transfer function $T_{z_\infty w}$ to be less than some specified levels. More precisely, the problem can be stated as follows.

For the system P, find an admissible controller K which satisfies the following design criteria:

- the closed-loop system must be asymptotically stable,
- the closed-loop transfer function from w to z_∞ satisfies the constraint:

$$\left\| T_{z_\infty w}(s) \right\|_\infty < \gamma, \tag{4}$$

  for a given real positive value γ,
- the closed-loop transfer function from w to z_2 must be minimized

$$\min \left\| T_{z_2 w}(s) \right\|_2 . \tag{5}$$

The task is to parameterize all suboptimal $\mathcal{H}_\infty$ dynamic controllers that stabilize the closed-loop system and satisfy the $\mathcal{H}_\infty$ constraint, and to find among them the controller that minimizes the standard $\mathcal{H}_2$ norm, [1, 3, 15].

Consider the closed-loop system in *Fig. 4*, which includes the suspension system and elements associated with the uncertainty models and performance objectives. Let the performance objectives be the heave acceleration of the sprung mass $\ddot{x}_1$, acceleration of the pitch angle $\ddot{\theta}$, suspension deflections $x_{1f} - x_{2f}$, and $x_{1r} - x_{2r}$, tire deflections of the unsprung masses x_{2f} and x_{2r}, and active forces generated by actuators u_f and u_r. Select the front and rear accelerations, x_{1f}, x_{1r}, of the body as the measured output. The weighting function $W_{p_1}, \ldots, W_{p_5}$ represent the different frequency domains of the performance outputs, namely the heave z_a and pitch z_θ acceleration, the suspension z_{sd} and tire z_{td} deflection and the control input z_u.

The measurement noises, which are denoted by n_f and n_r are taken into consideration in the design process. The random disturbances are denoted by w_f and w_r. Because of the effects of the external signals on the system, weighting functions W_w and W_n are applied to the disturbances and noises. The unmodelled dynamics is represented by W_R and Δ_M. It is assumed that the transfer function W_R is known, and it reflects the uncertainty in the model. The transfer function Δ_M is assumed to be stable and unknown with the norm condition, $\|\Delta_M\|_\infty < 1$. In the diagram, e is the input of the perturbation, d is its output. The augmented system to be controlled is shown in *Fig. 4*.

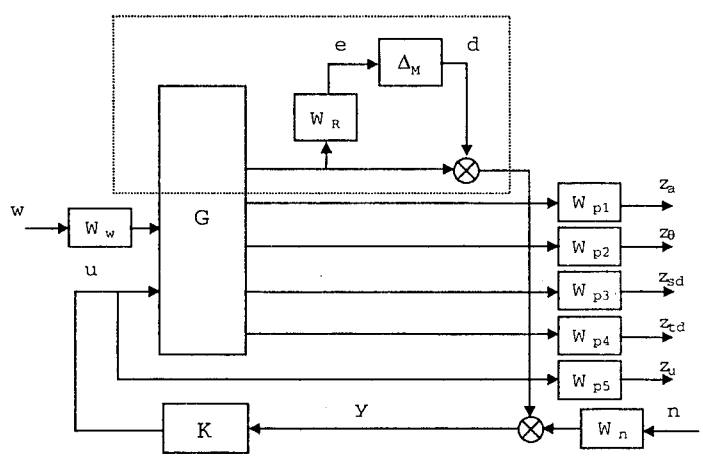


Fig. 4. The closed-loop interconnection structure

As a consequence, the $\mathcal{H}_2$ performance outputs and the $\mathcal{H}_\infty$ performance outputs are the following:

$$z_2 = \begin{bmatrix} z_a & z_\theta & z_{sd} & z_{td} & z_u \end{bmatrix}^T \tag{6}$$

$$z_\infty = \begin{bmatrix} e & z_a & z_\theta \end{bmatrix}^T . \tag{7}$$

Now, the design setup in *Fig. 4* must be formalized as a standard design problem, as illustrated in *Fig. 5*.

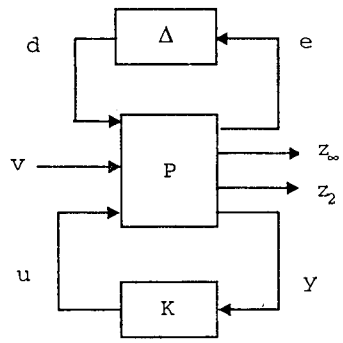


Fig. 5. The (P,K) structure with uncertainty

By applying the weighting functions and the compensator, the augmented plant and the closed-loop system can be formalized as the following forms:

$$\begin{bmatrix} e \\ z_\infty \\ z_2 \\ y \end{bmatrix} = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \begin{bmatrix} v_d \\ u \end{bmatrix} \tag{8}$$

$$\begin{bmatrix} e \\ z_\infty \\ z_2 \end{bmatrix} = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \\ M_{31} & M_{32} \end{bmatrix} \begin{bmatrix} d \\ v \end{bmatrix}, \tag{9}$$

where

$$v_d = \begin{bmatrix} d & w & n \end{bmatrix}^T$$
$$v = \begin{bmatrix} w & n \end{bmatrix}^T.$$

4. Demonstration Example

The nominal parameters which are used in the design procedure are in the Table 1 in the Appendix. In the first step of the control design, the uncertainty weighting function W_R and the performance weighting function W_P must be selected. It is assumed that in the low frequency domain disturbances at the sprung mass heave and pitch accelerations should be rejected by a factor of 0.5 by using $W_{p_1} = W_{p_2} = 0.2\frac{S+50}{S+200}$, and at the suspension deflection by a factor of 1 by using

$W_{p_3} = \text{diag}\left[0.029\frac{S+350}{S+10}, 0.029\frac{S+350}{S+10}\right]$ for front and rear suspension, respectively. Let the frequency weighting function for the tire deflection be $W_{p_4} = \text{diag}\,[1, 1]$, for the control force $W_{p_5} = \text{diag}\left[4 \cdot 10^{-3}, 4 \cdot 10^{-3}\right]$. It is assumed that the sensor noise is 0.001 m/s^2 at the front and rear body acceleration in the whole frequency domain. The weighting function of the unmodelled dynamics is selected as follows: $W_R = 1.875\frac{S+2}{S+25}$. The weighting functions are illustrated in *Fig. 6*.

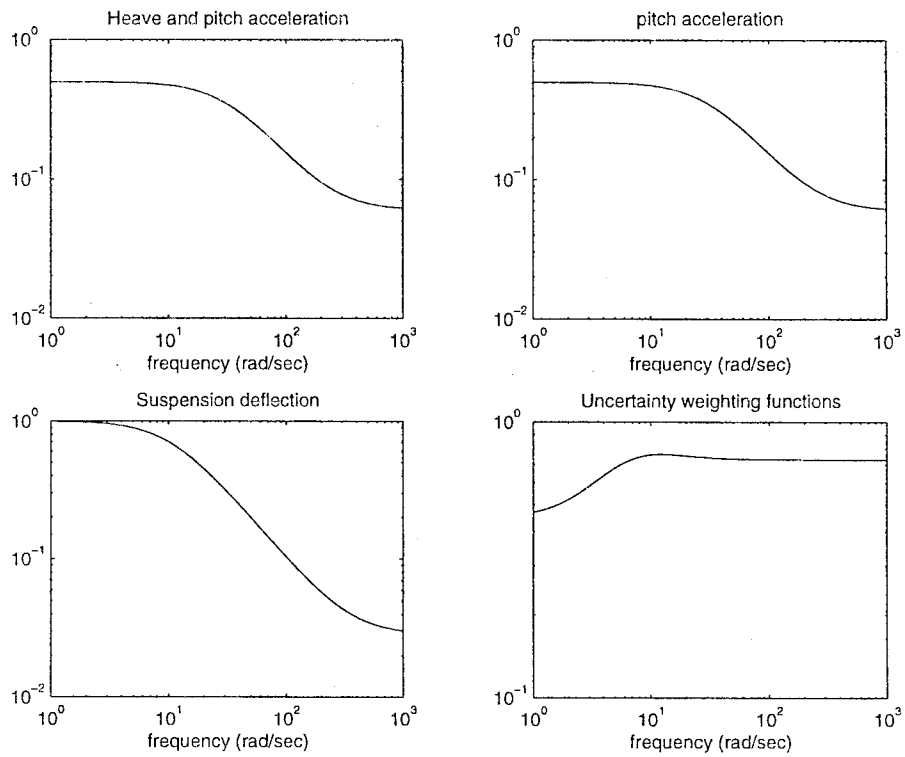


Fig. 6. Uncertainty and performance weighting functions.

In the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control design, a balance must be set up between the $\mathcal{H}_2$ norm and $\mathcal{H}_\infty$ norm, i.e. between nominal performance and robust stability specifications. In order to create a balance between the $\mathcal{H}_2$ and $\mathcal{H}_\infty$, the $\mathcal{H}_\infty$ versus $\mathcal{H}_2$ curve is analyzed. When modifying γ from a relatively high value, in each case the $\mathcal{H}_\infty$ norm and the $\mathcal{H}_2$ norm must be estimated to plot the points of the curve.

A gamma value that reduces the $\mathcal{H}_\infty$ norm significantly with little increase in the $\mathcal{H}_2$ norm should be selected, thus the selected γ gives an appropriate balance to the two norms. Due to the subjectivity of this selection, several controllers are constructed using some acceptable γ values. In these cases, the closed loop transfer functions between the disturbance w and performance outputs z_2 and z_∞ are com-

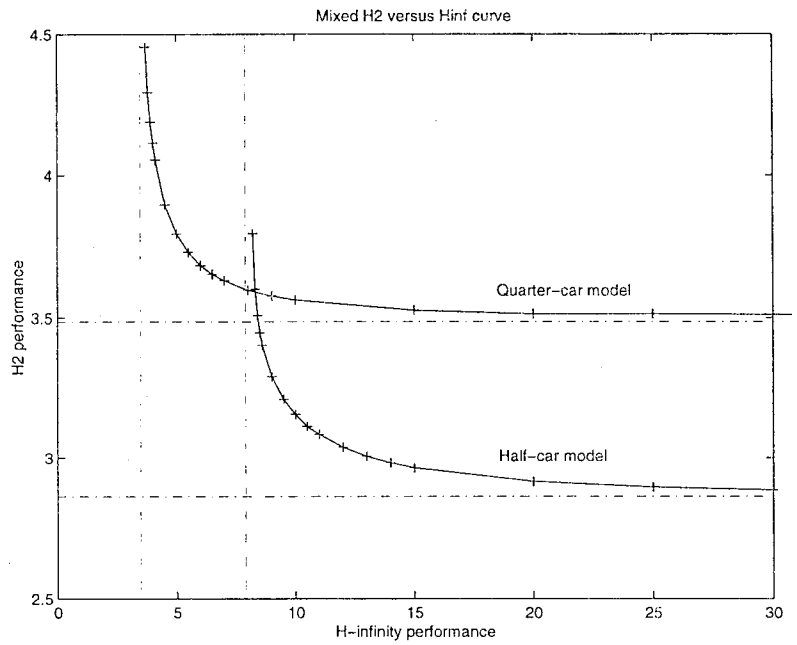
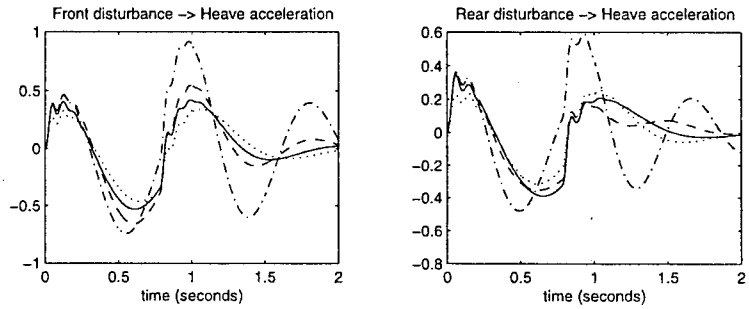


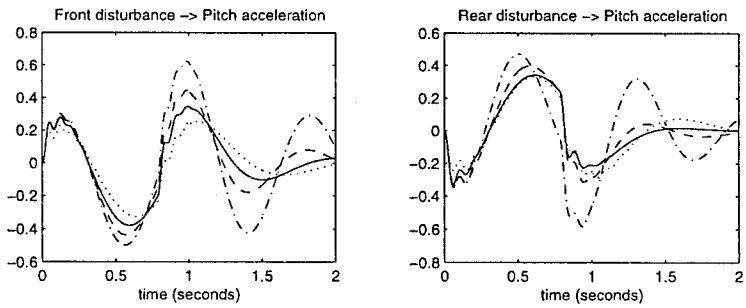
Fig. 7. Mixed $\mathcal{H}_2$ and $\mathcal{H}_\infty$ performances in different model structures.

puted to plot their time and frequency responses. In order to support the selection of γ, the pure $\mathcal{H}_2$ compensator design is performed to examine the closed loop between w and z_2. Moreover, the pure $\mathcal{H}_\infty$ controller is also designed to examine the closed loop between w and z_∞. This is a heuristic comparison possibility between the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ compensators. *Fig. 7* shows the mixed $\mathcal{H}_2$ versus $\mathcal{H}_\infty$ curves by using different suspension structures, i.e. the quarter-car and the half-car models.

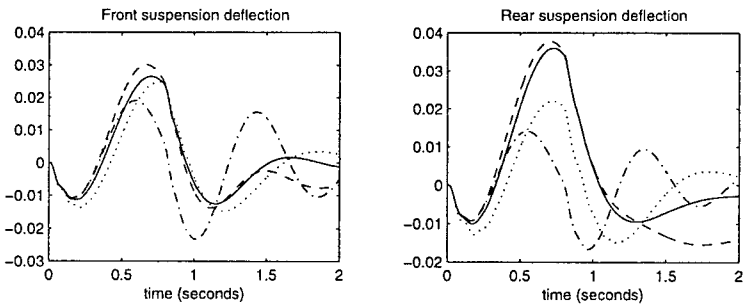
The effects of the disturbance on the body mass acceleration, the suspension, tire deflection, on the control force are illustrated in the time domain in *Fig. 8*. In the example, the input signal is simulated as a bump with 0.02 m maximal value. The solid line corresponds to the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ synthesis, the dashed line to the $\mathcal{H}_\infty$ synthesis, the dotted line to the LQG design, and the dashed-dotted line to the passive system. The frequency responses are shown in*Fig. 9*.



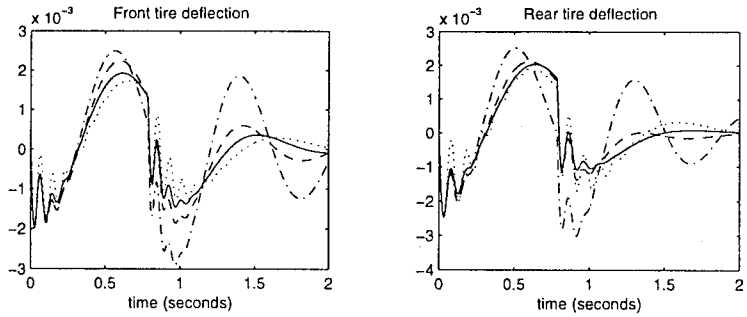
(a) Heave accelerations



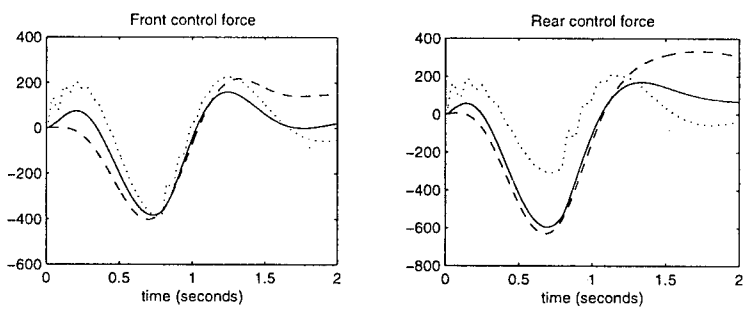
(b) Pitch accelerations



(c) Suspension deflections



(d) Tire deflections

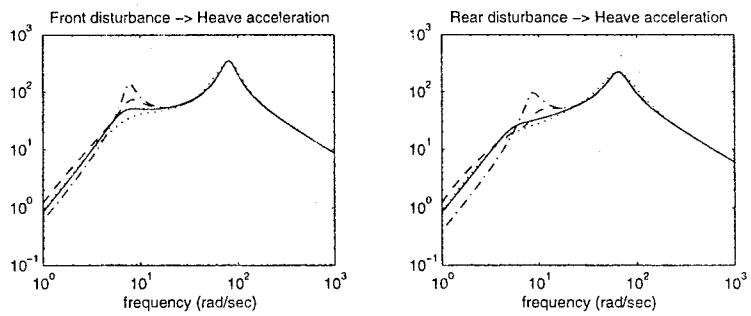


(e) Control forces

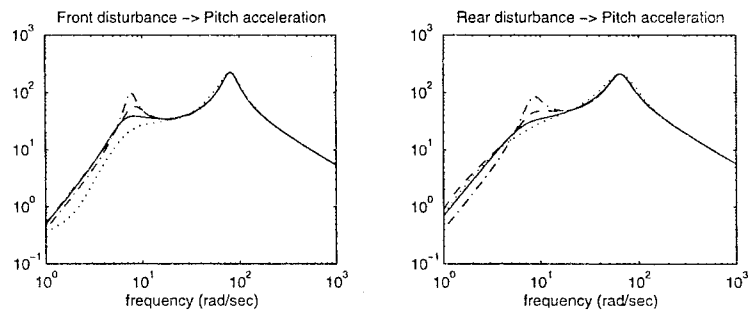
Fig. 8. Time responses of the controlled system

5. Conclusions

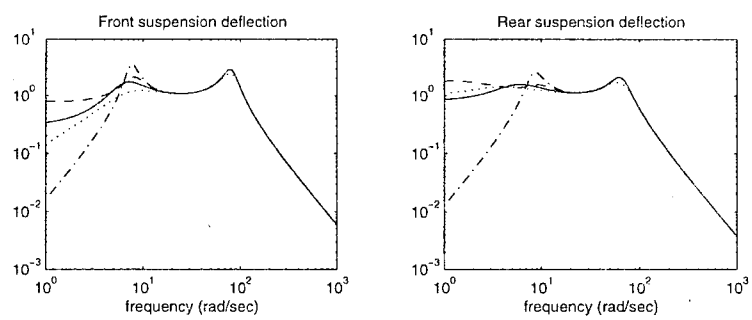
This paper has presented the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control design to solve the active suspension problem. The advantage of this method is that it provides excellent disturbance rejection as a result of the $\mathcal{H}_2$ criterion, and a good performance and stability margin as a result of the $\mathcal{H}_\infty$ criterion. The demonstration example has illustrated that this method is well suited to controllers for active suspension since the technique generates a set of controllers that balance between nominal performance and robust stability requirements.



(a) Heave accelerations



(b) Pitch accelerations



(c) Suspension deflections

Fig. 9. Frequency responses of the controlled system

Acknowledgement

This work was supported by the Hungarian National Science Foundation (OTKA) under the grant #T030182 which is gratefully acknowledged. Dr. Gáspár was supported by the

Bolyai scholarship granted by the Hungarian Academy of Sciences, which is gratefully acknowledged.

References

[1] BERNSTEIN, D. S.–HADDAD, W.M.: LQG Control with an $\mathcal{H}_\infty$ Performance Bound, *IEEE Trans. on Automatic Control*, **34**, pp. 293–305. (1989).

[2] CROLLA, D. A.–ABDEL HADY, M. B. A.: Active Suspension control; Performance Comparisons Using Control Laws Applied to a Full Vehicle Model, *Vehicle System Dynamics*, **20**, pp. 107–120. (1991).

[3] DOYLE, J.– ZHOU, K.–GLOVER, K.–BODENHEIMER, B.: Mixed $\mathcal{H}_2$ and $\mathcal{H}_\infty$ performance objectives. II. Optimal control, *IEEE Trans. on Automatic Control*, **39**, pp. 1575–1587. (1994).

[4] HROVAT, D.: Optimal Active Suspension Structures for Quarter-car Vehicle Models, *Automatica*, **26**, pp. 845–860., (1990)

[5] JAGER, B. B.: Multiobjective Control: An Overview, *Proc. of the 36th Conf. on Decision and Control*, pp. 440–445. (1997).

[6] MICHELBERGER, P.–BOKOR, J.–PALKOVICS, L.: Robust Design of Active Suspension System, *International Journal of Vehicle Design*, **14**, pp. 145–165. (1994).

[7] MORAN, A.–NAGAI, M.: Performance Analysis of Vehicle Active Suspension with Hμ Robust Control, *Proc. of the 1st Int. Conf. on Motion and Vibration Control*, Yokohama, pp. 756–761., (1992)

[8] PALKOVICS, L.–GASPAR,P.–BOKOR, J.: Design of Active Suspension System in the Presence of Physical Parameter Uncertainties, *Proc. American Control Conf.*, **1**, pp. 696-700. (1993).

[9] SHARP, R. S.–CROLLA, D. A.: Road Vehicle Suspension System Design – a Review, *Vehicle System Dynamics*, **16**, pp. 167-192., (1987)

[10] SCHERER, C.: Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ Control, *In: A. Isidori (ed.), Trends in Control: A European Perspective, Springer, Berlin*, pp. 173-216. (1995).

[11] RIDGELY, D. B.–VALAVANI, L.–DAHLEH, M.–STEIN, G.: Solution to the General Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ Control Problem – Necessary Conditions for Optimality, *Proc. American Control Conf.*, 2, pp. 1348–1357. (1992).

[12] SEINFELD, D.R. –HADDAD, W. M.– BERNSTEIN,D.S.– NETT, C. N.: $\mathcal{H}_2/\mathcal{H}_\infty$ Controller Synthesis: Illustrative Numerical Results Via Quasi-Newton Methods, *Proc. American Control Conf.*, pp. 1155–1156. (1990).

[13] STEINBUCH, M.–BOSGRA, O. H.: Static and Fixed Order Dynamic Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ Optimal Control, *Proc. American Control Conf.*, pp. 1137–1142. (1991).

[14] YAMASHITA, M.–FUJIMORI, K.–HAYAKAWA, K.–KIMURA, H.: Application of $\mathcal{H}_\infty$ control to active suspension system, *Automatica*, **30**, pp. 1717-1729., (1994)

[15] YEH, H.H.– S.S. BANDA, – B.C. CHANG,: Necessary and Sufficient Conditions for Mixed $\mathcal{H}_2$ and $\mathcal{H}_\infty$ Optimal Control, *IEEE Trans. on Automatic Control*, **37**, pp. 355-358. (1992).

Appendix

Table 1. Parameters of the half car model

Parameters (symbols)	Value
sprung mass (m_s)	580 kg
front unsprung mass (m_{af})	40 kg
rear unsprung mass (m_{ar})	40 kg
front suspension stiffness (k_{sf})	23500 N/m
rear suspension stiffness (k_{sr})	25500 N/m
front tire stiffness (k_{tf})	19900 N/m
rear tire stiffness (k_{tr})	19900 N/m
front suspension damping (b_{sf})	1000 N/m/s
rear suspension damping (b_{sr})	1100 N/m/s

PERIODICA POLYTECHNICA SER. TRANSP. ENG. VOL. 28, NO. 1–2, PP. 17–22 (2000)

THE CONDITIONS OF THE APPLICATION OF ETCS IN HUNGARY

Géza TARNAI*, Balázs SÁGHI* and László MOSÓCZI**

*Department of Transport Automation
Budapest University of Technology and Economics
H–1521 Budapest, Hungary
Tel: +36 1 463 1013
e-mail: {tarnai|saghi}@kaut.kka.bme.hu
**Hungarian State Railways
H–1062 Budapest, Andrássy út 73-75, Hungary
Tel.: +36 1 432 3390
e-mail: MosocziL@mavrt.hu.

Received: October 5, 2000

Abstract

The realization of the interoperability between the partners of an intermodal transportation chain as well as inside a single transportation mode heavily influences the quality of the whole transportation process. The railway can only be a competitive partner of the intermodal transportation chain, if the railway network is equipped by homogeneous infrastructure through several countries. Train Control Systems play here an important role. The main goal of this paper is to introduce the efforts for a unified European Train Control System (ETCS) and to present the results achieved up to now.

Keywords: train control, railway signalling, ETCS.

1. ETCS in Europe

1.1. Goals of the Introduction of ETCS

For the sake of stopping the loosing ground of the railways, the European Union (EU) adopted as a first step the aim of improving the competitiveness of railways in the member states. It was decided that the main economic and political centres should be interconnected with a passenger transport system offering fast travel, the ratio of combined cargo transport should be increased, the railway technical standards should be unified (facilitating interoperability) and at the border crossing stations the administration and staying time should be reduced to a minimum.

Over the years, a number of different Automatic Train Control (ATC) systems have been introduced in different countries at different times. These systems are incompatible with each other. Only a few are used in more than one country, and in cases where the same basic system has been adopted in different countries, there have been differences in detailed development which has resulted in incompatible versions. The large number of different national systems in use is a major obstacle for cross-border traffic in Europe.

According to the principles outlined above, in the early nineties, on recommendation by the EU, the development of the unified European Train Control System constituting a part of the unified European Railway Traffic Management System (ERTMS) was started.

There are several factors that force the introduction of ETCS. The needs can be classified as economic, business and technical aspects.

Economic aspects

The tendency of liberalization in the transport world means that in future the railways are going to be more competitive. The system of the future, therefore, is intended not only to improve operating efficiency, as will be detailed below, but also to seize every opportunity to reduce acquisition and maintenance costs. The adoption of a uniform system at European level will result in cost reductions, as system development and maintenance can be carried out in a uniform manner. Furthermore, the fact that all the railways will be using the same system means that the system components can be produced in larger quantities.

Business needs

All railways have the same business need: to be able to provide market-oriented, reliable, safe and cost-effective train transportation for both passengers and freight. One of the prerequisites for this is to have an appropriate system for automatic train control (ATC) and protection (ATP). The business aims of the railways can be listed as follows:

- enhance the safety of the train traffic by equipping lines with ATC/ATP,
- increase the carrying capacity of the line by taking the headway and speed to the limits physically possible, while keeping the existing safety level.

The current technical solutions for intentional traffic do not meet the railways' long-term reliability and efficiency requirements because the train can pass the borders only with the changing of the traction units due to the different ATC systems.

Technological needs

Increased use of microprocessors in conjunction with electronic data processing has led to the major changes in the systems, which control and protect train traffic. More and more significant parts of the system such as points or signals will be computer controlled. Such computers may be equipped on the ground or on the board of the vehicles and will be connected by means of data transmission system. The increasing complexity of the train control systems means that the cost of developing

new ones is continuously increasing. Smaller railways and manufacturers are less and less able to cover such costs. Joint work on the development of an ETCS system for the future will produce synergic effects for the railways and industry as a result of a sensible division of responsibility.

1.2. Benefits of Installing ETCS

The benefits of installing ETCS on the European railway network can be summarized as follows:

- Faster and better international traffic, it will become possible to apply more reliable and more cost-effective equipment to rolling stock used for high-speed passenger traffic.
- Reduced headways on highly loaded parts of the network, additional level of functionality, available to all the railways, such as driving on electronic sight, will allow those parts of the network which present bottlenecks to be loaded to their physical limit.
- Reduction in costs of building and renovating lines and stations and of replacing outdated ATC systems.
- Reduction in a workload for the technical departments of the individual railways, as a result of the provision of uniform documents and aids for the construction and maintenance of the equipment, drawn up in accordance with the European standards.
- The establishment of a competition among the manufacturers.
- The precondition for the future harmonization in other areas of operation control.
- Increased availability, based on the evaluation of EU-wide failure statistics.

Obviously, the significance of these benefits can be different for each railway.

1.3. ETCS Application Levels

ETCS is defined to be a very flexible system offering easy configuration features to different application levels. Fall-back solutions and interoperability with existing national ATP systems in Europe are provided. The level to which a part of a line or the vehicles are equipped with ETCS equipment determines the level of performance of the system, i.e. the number and sophistication of the functions available on that line. On the other hand, each level of application also defines the amount of equipment to be installed. The application levels are defined as follows.

Level 1 provides Automatic Train Protection with supervision of line speed and braking to restrictive targets. It is proposed to be used in connection with fixed blocks, signals, and train detection functions outside the ETCS system (e.g. by track circuits or axle counters). Level 1 can be implemented with spot transmission

(balises, short loops) or with target information updating (infill) by medium length loops or radio.

The higher application levels provide Automatic Train Control (ATC) with conventional train detection and fixed block sections with or without lineside signals (Level 2), or with continuous train location, radio block, and possibility for moving block operation (Level 3).

2. Introduction of ETCS into the Hungarian Railway Network

2.1. ETCS-VB Project

In 1993 the Hungarian State Railways (MÁV) decided on starting the introduction on its network of a new, point-type (intermittent) train control system in relation to the elevated train speed. Having studied the market and the international development trends, tendencies, the management of the company decided not to implement an already existing train control system on the network but rather to install the ETCS Level 1. This was justified by the fact that owing to the intent of joining the EU, it is expedient to choose it, as this system is advantageous for achieving the interoperability of the main lines and because by the middle of the nineties the development of the system has advanced [4].

Accepting the proposal of the EU and the International Union of Railways (UIC), the Austrian Federal Railways (ÖBB) and the Hungarian State Railways (MÁV) agreed that with certain conditions they will install the ETCS system on the Budapest – Vienna line.

Following a preparatory period of nearly one and a half year, the ETCS-VB (European Train Control System Vienna – Budapest) Consortium has been established by ÖBB, MÁV, five industrial companies and the ARSENAL Research and Development Centre (Vienna). The Consortium has won the support of EU. Beside the Spanish, Italian, German and French projects, the ETCS-VB is the only project with the participation of an East-European country and the only one incorporating a border crossing.

The tasks to be solved in the ETCS-VB pilot implementation project are illustrated in *Fig. 1*. The project (large gray box) is a part of the overall ERTMS/ETCS project and surrounded by other ETCS specifications, definitions and test projects [2].

2.2. Aims of the Introduction of ETCS in Hungary

The setting up of the experimental section including the Hungarian – Austrian border crossing and the implementation of the test program was planned with the main aims as follows:

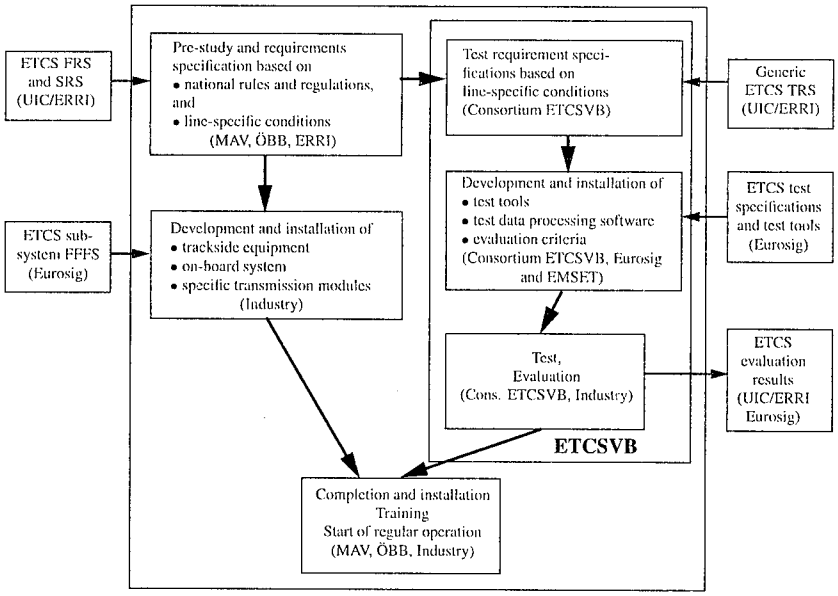


Fig. 1. ETCS pilot implementation project Vienna – Budapest (task overview) (ERRI: European Railway Research Institute, UIC: Union Internationale des Chemins de Fer, EUROSIG: Consortium of European Manufacturers, FRS: Functional Requirement Specification, SRS: System Requirement Specification)

- investigation of installation of the system on a real railway track section without interrupting the scheduled traffic,
- testing the possibility of interfacing to the existing safety equipment and the feasibility of fulfilling the prescribed functions,
- feasibility of compliance with the user requirements formulated by the railway companies, and
- checking the possibility of solving the technical questions of border crossing.

3. Evaluation of the ETCS VB Project

The pilot plant organized by the MÁV (on the Hungarian part) has proven that the specified system is viable and ready for operation [3].

In the course of the tests, the basic functions of the lineside signal transmitter elements were checked in regular operation and also in a simulated breakdown. It has been proven that the system is capable of triggering a safety reaction defined by the railways in case of a breakdown endangering the safe operation that leads finally to the stopping of the train.

The tests have clearly demonstrated that the guidelines formulated by the European Union are basically correct. The system corresponding to this can be fitted to the existing station and lineside safety equipment at the railway lines pertaining to the international basic network (such as the Budapest–Hegyeshalom line section).

Furthermore it has been proven that the system complies with the conditions of high-speed traffic, even more in some cases it can contribute to increasing the throughput capacity of the line, less traffic disturbing in the case of a possible failure of the lineside equipment.

For MÁV the participation in the first experimental project of ETCS was an advantage with particular attention to formulating the demands and expectations related to the ETCS system and to the further installation of the system along the entire Budapest–Hegyeshalom section.

For the better presentation of the lineside and on-board systems a simulation and training system introducing the functionality of ETCS has been developed at the TU Budapest, Department of Transport Automation. The system is capable of being applied for the primary training of the engine drivers and the persons responsible for the maintenance of the lineside equipment, and at the same time for providing information of sufficient quantity and level for the decision-makers [1].

The pilot plant – considering also the preparatory, design, implementation and evaluation stages – has shown the areas where new regulations should be worked out and new technology provisions should be introduced.

References

[1] ARANYOSY, Z. – MOSÓCZI, L. – RÁCZ, G. – TARNAI, G.: Training the Personnel Using Simulator and Training System. *World Congress on Railway Research Conference*. Florence, Italy, 1997. pp. 745–751.

[2] DOPPELBAUER, J. – RHEIN, D. – MOSÓCZI, L. – ZIERL, A.: The New European Train Control System (ETCS): Installation Problems on an Existing Line. *World Congress on Railway Research Conference*. Colorado Springs, USA, 1996, pp. 151–157.

[3] ETCSVB Project. Final Report for Publication. 29.02.2000. p. 106.

[4] GÁL, I. – MOSÓCZI, L.: Development of the Signalling System and Introduction of ERTMS/ETCS at Hungarian State Railways. *Railway Signalling and Telecommunication Technology Conference*. Tokyo, Japan, 1998, pp. 235–241.

PERIODICA POLYTECHNICA SER. TRANSP. ENG. VOL. 28, NO. 1–2, PP. 23–38 (2000)

THE IMPACT OF TELECOMMUNICATION AND TRANSPORT ON SPATIAL BEHAVIOUR

Dirk ZUMKELLER

Institute for Transport Studies
University of Karlsruhe (TH)
2000 Karlsruhe, Germany

Received: Nov. 10, 2000

Abstract

Telecommunication has not only for the sender but as well for the addressee both mobile and immobile elements. Regarding telecommunication in the interpersonal context with the related traffic behaviour, it becomes clear that telecommunication has so far an unknown influence on our spatial behaviour.

Based on these considerations a concept is being developed to demonstrate the influence, the use and the increasing penetration of communication and information media on spatial behaviour of humans. The question, which effects are to be considered from this for the future and in which way it affects planning of interventions in the traffic sector, follows directly.

On the basis of empirical results from Germany, Sweden and Korea it is shown that additional communication and information possibilities have no decreasing affect on the physical mobility of humans. Based on this result the advantages of novel communication and information services are being systematised to analyse the impacts on spatial behaviour in detail. For this it is possible to fall back on data-sets ranging from the mega-city Seoul over cities and rural regions in Germany to remote areas in Sweden.

So it is to be expected that certain time-consuming, standardise and according to their nature suitable activities/ trips (e. g. telebanking) might be substituted in the every day live. At the same time, however, it is to be expected that far distant destinations can be more easily investigated by better information and communication possibilities, in order to lead afterwards to additional physical mobility – thus an induction of physical transport appears this way. The increases are to be expected fewer in everyday life transport, since the financial and temporal budget restrictions are effective here due to capacity limitations of the traffic system. Rather increases in the weekend and holiday traffic are to be expected, where either by the generated interest via simplified information access or by the decrease of initial trave) thresholds (reduction of uncertainties concerning the selected destination by telecommunications) additional journeys can be performed. Within the leisure area and the global business and service area activities and appropriate journeys are thus generated, which would not have been possible without existence of the electronic media.

Finally, it is stated, which (feedback-) effects result on the structure of demand, if more spontaneous acts caused by information and communication technologies provoke critical and on a long-term basis not calculable effects.

Keywords: behavior, behavioral, complementarity, coordination, interdependencies, intrapersonal, labor, online, superposition, teleactivities, telecom, teleservice, telework.

1. The Extension of Spatial Action Fields

Telecommunication has not only for the sender but also for the addressee both mobile and immobile elements. The immobile element can be understood as the

fact not to be forced to move physically whereas the mobile element applies both to the sender and the addressee since the communicated information extends their spatial action field. Thus the only distinction between sender and addressee is the question whether the one or the other was active or passive.

When we try to translate this concept to physical transport we start to understand why the passive element of transport (to be visited) is much less important than the active element (to move), since the associated cost in terms of time, money, energy etc. is much higher for the active part. Nevertheless, we have to state that the passive part of transport contributes to the extension of the spatial action field of the visited person in the same way as telecommunication does.

These very basic definitions indicate a strong interrelation between transport and telecommunication both contributing in a more or less complementary manner to the (never-ending?) extension of spatial behaviour of humans [1, 2].

First quantitative but descriptive results concerning the interrelationship between physical and virtual transport (telecommunication) become possible by expanding our observation of spatial behaviour patterns by including the dimension of overcoming space on a virtual level [3].

The results of this empirical approach have clearly indicated that the interrelationship goes far beyond what is currently understood and incorporated in our general understanding. A basic phenomenon, however, has already emerged: there is a further increase to be expected in the field of telecommunication as well as in the field of transport giving the background of increasing spatial division of labour and further dispersion and specialization of private life. However, this will happen consistently, interrelated between transport and communication, probably with a stronger growth in the fields of telecommunication due to the lower cost in terms of personal time, cost, pollution etc.

And since it has been shown that evidently there is little hope for substitution theories and hopes for decreasing transport volumes it can be assumed that human positive curiosity is the driving factor for the demand to extend spatial action fields as much as possible – if not restricted by personal intellectual, time and cost budgets or social and environmental limitations (*Fig. 1*).

1.1. Some Coherent Spatial Elements of Transport and Telecommunication

Having in mind that the penetration of the telephone was strongly correlated with the growth of the transport market [4], it is postulated that – among other causes – new telecommunication services can be understood as major driving factors to overcome spatial separation. Thus, the hypothesis is that new telecommunication services can cause more transport since they ease the investigation of new (distant) destinations.

The following empirical results (*Fig. 2 – 5*) which are derived from an integrated transport and telecommunication survey (see chapter 2.1) in the Baden-Württemberg region [5] shall illustrate the spatial and temporal co-ordination of

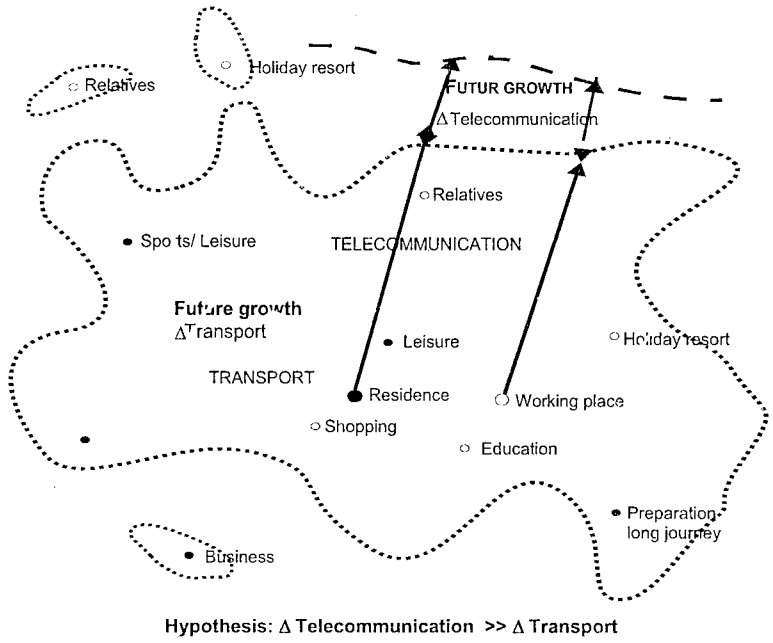


Fig. 1. Spatial action fields for traffic and communication

the interpersonal use of modes and media. Particularly, the differences in the use of media between certain person groups in a rural or urban context (*Fig. 4*) indicate the coverage of much more distant destinations by urban people – especially those being equipped on a high level. The same pattern appears when looking at all destinations of all trips and contacts originating from a small town (Weinberg), situated in a 50 km catchment area of Stuttgart (*Fig. 5*).

1.2. Availability, Competence and Acceptance of Modes and Media

Availability of media and competence of their use vary with age. Particularly very young and old age groups develop rather different lifestyles resulting in an acceptance of media ranging from fashion to rejection [6]. *Figs. 6* and *7* clearly indicate this phenomenon at home and at work (affecting only one half of the population).

Obviously, this basic situation affects the acceptance of media and their accelerating change. In 1998, 15% of all Germans had access to the internet and online services; more than 9% used it regularly, 60% of them were between 20 and 39 years. Little fantasy is necessary to imagine that in 10 years this picture will change dramatically both in size and shape

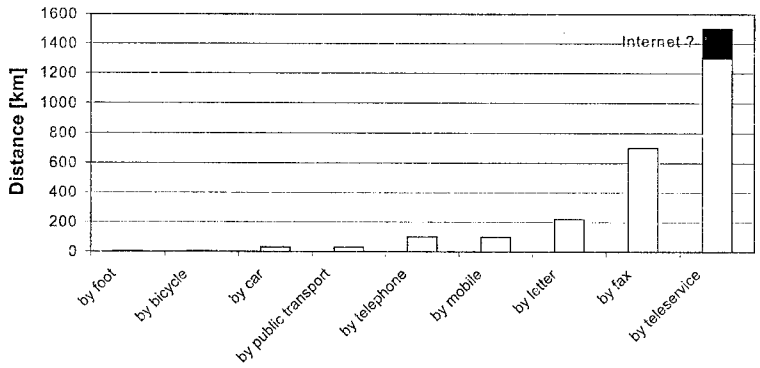


Fig. 2. Spatial elements of modes and media

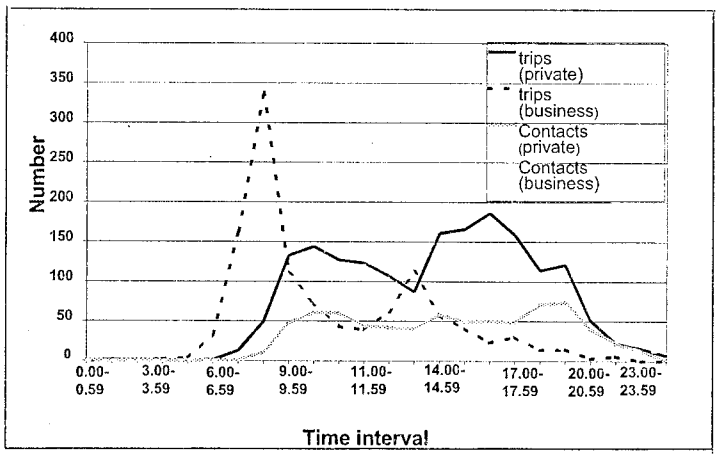


Fig. 3. Temporal distribution of traffic or telecommunications activities

2. Substitution versus Complementarity

The certainly interesting question of mutual interdependences between transport and telecommunication cannot be answered finally based on results of already completed surveys. The most eminent role of telecommunication as a tool to plan and perform physical transport became visible as well as the option to perform trips more efficiently. This can be associated with the substitution of more or less standardized trips. But, on the other hand, telecommunication opens the window to get in contact with far away destinations possibly resulting in a related long-distance travel. This view is consistent with the historical fact that transport and development of media were more or less correlated.

Type of area		Persons with mobile		Persons with teleservice		All Persons	
		Number per person per day	average distance [km]	Number per person per day	average distance [km]	Number per person per day	average distance [km]
Rural area	Trips	3.4	16.9	3.7	12.6	4.3	11.4
	Contacts	7.3	49.0	6.1	47.0	3.4	87.0
Urban area	Trips	5.6	9.7	4.5	11.1	4.0	9.0
	Contacts	6.0	172.0	3.6	211.0	2.9	162.0
Total	Trips	4.8	11.7	4.2	11.6	4.1	9.9
	Contacts	6.5	139.0	4.4	167.0	3.1	140.0

Fig. 4. Spatial separation overcoming behaviour as a function of the communication equipment of the households

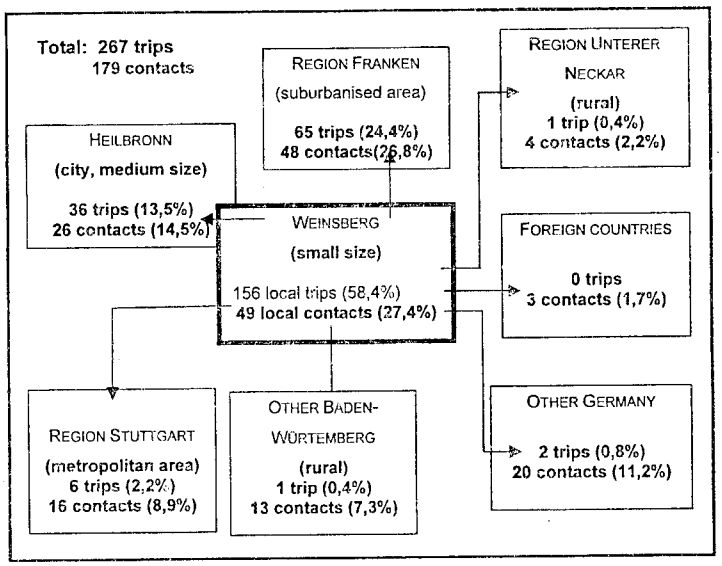


Fig. 5. Destination (trips/contacts) of interviewed persons in Weinsberg

Particularly a glance at lifestyle types shows that new telecommunication applications have a significantly higher use by young, male, highly qualified and mobile people. Thus we have to expect that only some decades later the picture may change totally. Is it not therefore necessary to observe our spatial behaviour in a new more comprehensive way?

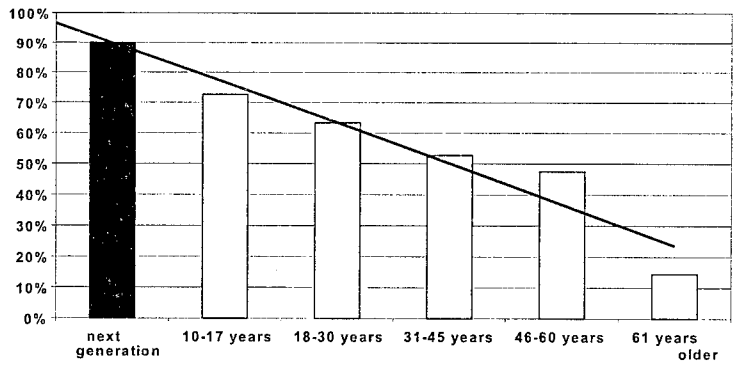


Fig. 6. PC-availability (at home) by age group (cohort effect)

Telecommunication service	(%)
Telephone	40.4
Fax machine	25.4
Cellular phone	5.8
PC	22.8
PC (with Internet access)	8.7
None	22.0
Other	1.0

Fig. 7. Telecommunication equipment at work

2.1. Observation of our Spatial Behaviour in a New Context

The aggregate traffic flows as well as the amount of bits and bytes exchanging locally are always the result of the n of individual behavioural patterns. When defining a new context of our spatial behaviour, priority is given to the observation of individual behaviour. Consequently, the individual has to form the elementary sampling unit of the following surveys, observing the spatial behaviour of humans in an interpersonal context of all trips and contacts.

Therefore an empirical investigation of the daily context of telecommunication and trips was carried out on a microscopic level, aiming at a deeper insight into the interpersonal frequency of using different modes and media (including cars, public transport, walking, cycling, telephone, fax, letters, mobile phones, e-mail, internet etc.) and their interrelationships [7].

Based on this empirical approach, meanwhile the following samples have been realized:

Bavaria	1994	2000 households	(5)
Baden-Württemberg	1997	750 households	(5)
Seoul (Korea)	1998	750 households	(8)
Sweden	since 1998	1000 households/ year	(9)

The results given in this paper are based on these surveys.

2.2. *Transport and Telecommunication Activity Patterns*

The integrated interpersonal observation of our spatial behaviour includes both physical transport and telecommunication forms of the empirical platform for potential model developments. Since the aim of such modeling approach is the estimation of impacts of new telecommunication services, the key issue of modeling is considered to be the combined pattern of physical and virtual off-home activities which are based on the following definitions:

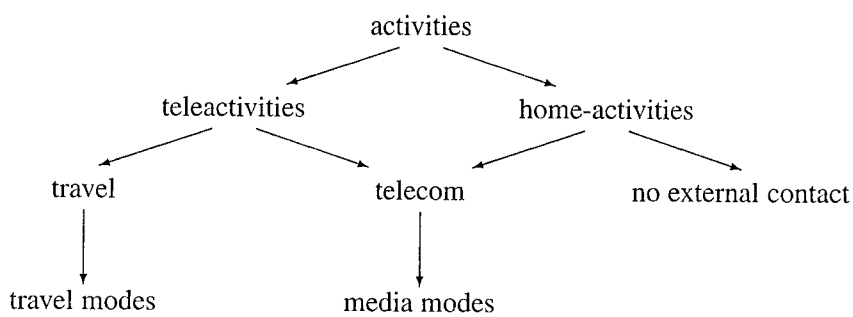


Fig. 8. Physical and virtual activities

Thus, an activity is described by the purpose and the spatial element, which could be constituted by one, two or more spatial elements (traffic zones etc.). The resulting activity pattern then forms a sequence of activities containing the information per activity already mentioned (*Fig. 9*).

activity pattern							
purpose	home		work		private contact		home
spatial element	residence		working place		person in zone x		residence
potential travel/communication choices							
choice A no travel	home		Telework		phone		home
choice B car travel	home	car	work	car	visit	car	home
choice C bus travel	home	bus	work	bus	phone		home

Fig. 9. Examples for combined travel/telecommunication activity patterns

For further simplification all activities are defined by the codes given in *Table 1*.

Table 1. Codes of activities

Code/ activity	Code/ activity
1 work	6 missing value
2 business	7 contact
3 education	8 leisure
4 shopping	9 private affairs
5 home	

Thus it follows that the inclusion of a combined trip/communication generation process is a necessary prerequisite for modeling interactions between modes and media. The basic characteristic of this module should be among others that

- either trips can be omitted due to contacts (substitution)
- or trips can be generated due to contacts (complementarity).

For further investigation the following form is selected in support of SCHULMEYERS work [9] for the representation of a trip-/contact sequence:

$$X - H - Y - 5 - Z$$

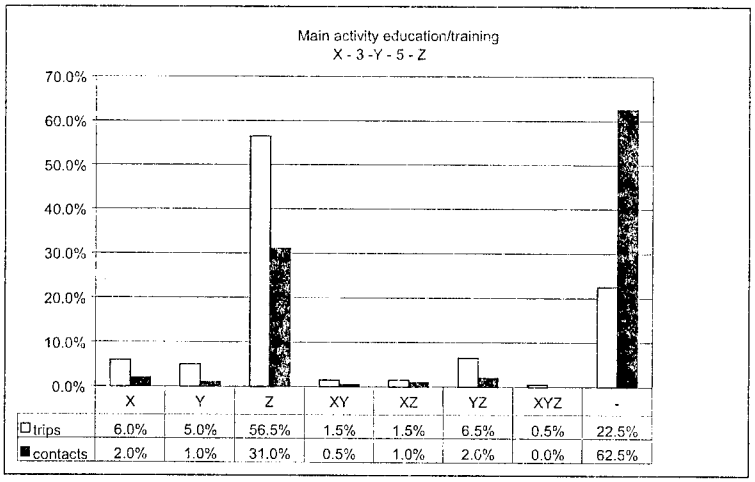
H: terminal characters for the main activity with $H \in \{1 \cap 2, 3, 4, 8 \cap 9\}$
5: terminal characters for the way home
X, Y, Z: nonterminal characters for sequence sections which stand

- before main activity (X)
- on the way to main activity and way home (Y)
- after the way home from main activity (Z)

Trips-/contact patterns	4	7	1	2	7	5	8	5
Formal representation	X		H	Y		5	Z	

In the example, the observed person goes shopping before going to work and executing a contact. Before she drives home from work, the person undertakes a business trip and executes a contact. After coming home from work she undertakes a leisure activity outside and subsequently drives home afterwards.

This form of illustration shall make it possible to examine trip-/contact patterns systematically, with the purpose to get pictures on structures of the combined trip/contact patterns [11]. These are classified by their main activities (education/ training, work, shopping, leisure). In the following, for the main activity 'education/ training' the frequence distributions of the individual sequence sections will



	X	Y	Z	XY	XZ	YZ	XYZ	-
trips	6.0%	5.0%	56.5%	1.5%	1.5%	6.5%	0.5%	22.5%
contacts	2.0%	1.0%	31.0%	0.5%	1.0%	2.0%	0.0%	62.5%

Fig. 10. Structure of trip-/contact patterns with the main activity education/training

be represented and commented, and they are distinguished by sequence sections containing trips or contacts.

In 56.5% of all investigated cases additional trips were made after returning home from a training or educational facility. 6.5% of all patterns included additional trips between these two locations. 22.5% of all patterns consisted of a direct trip to the educational/training facility plus a corresponding way home only. Additionally, contacts were transacted mostly after returning home. However, in 62.5% of all cases no contacts were transacted at all. Therefore, the typical structure of a trip/contact pattern with the main activity education/training is as follows: 3 – 5 – Z. Subsequently, the structure of the sequence section Z is to be examined more detailed. Thus, the following table contains selected sequence sections, occurring together with the main activity education/training.

This table illustrates that a leisure activity, combined with a way home is the most frequent basis sequence[1]. Furthermore, this basis sequence is part of most sequence sections. In addition, after the main activity the leisure activity is dominant for the individual behaviour of a person in the formation. 6% of all sequence sections contain another education trip. That means, the observed person drives home for lunch and returns to training facility afterwards.

Additionally, it has to be stated that still 48 different sequence sections can be determined despite the high aggregation level.

The table also shows that trips are often combined with contacts and 10.6% of all sequence sections consist of contacts only. As mentioned before it is useful to localize the individual contacts in a combined sequence section for an evaluation of

[1]Basics sequences are defined to consist of one activity and a trip back home only.

Table 2. Selected sequence sections (3-5-Z)

Sequence section Z	Absolute frequency	Relative frequency	Explanation
8 – 5	28	21.2%	Leisure time, home
7 – 8 – 5	18	13.6%	Contact, leisure time, home
7	14	10.6%	Contact
3 – 5	8	6.0%	Education/training, home
8 – 5 – 8 – 5	5	3.8%	Leisure time, home, leisure time
9 – 5	3	2.3%	Private transactions, home
8 – 5 – 7	3	2.3%	Leisure time, home, contact
7 – 3 – 5	3	2.3%	Contact, education/training, home
4 – 5 – 8 – 5	3	2.3%	Shopping, home, leisure time, home
7 – 8 – 5 – 7	3	2.3%	Contact, leisure time, home, contact
	12	9.1%	Twice occurring sequence sections
	32	24.2%	Once occurring sequence sections

the effect of telecommunication on the individual traffic behaviour. *Table 3* shows the position of contacts in sequence section Z.

Table 3. Position of contacts in the sequence section Z (3-5-Z)

Position of contacts	Absolute frequency	Relative frequency
'only contacts'	14	19.7%
before a trip	41	57.8%
between two trips	1	1.4%
before a way home	2	2.8%
at the end of the sequence	13	18.3%

19.7% of the sequence sections contain contacts only. 57.8% of contacts in sequence section Z are performed before a trip. To do predictions on the physical relevance of these contacts one has to know why these contacts are examined (*Table 4*). But strengthened predicates cannot be found, because the specification of the contact purpose often becomes unimportant and specifications like 'chat' or 'private' do not permit conclusions concerning the relevance. Nevertheless, here shall be assumed that contacts can be useful for following activities and therefore could represent the preparation of physical traffic, too.

The following *Fig. 11* gives a first glance at the main activity 'work' to show the existing differences (for further main activities see [11]).

Table 4. Contact purposes

Contact purpose	Proportion [%]
private (without specification)	12.6
make appointment	12.2
chat	12.4
pass on or receive an information	12.1
advertisement/shopping/order	3.4
telebanking	2.4
other	9.4
no specification	35.5

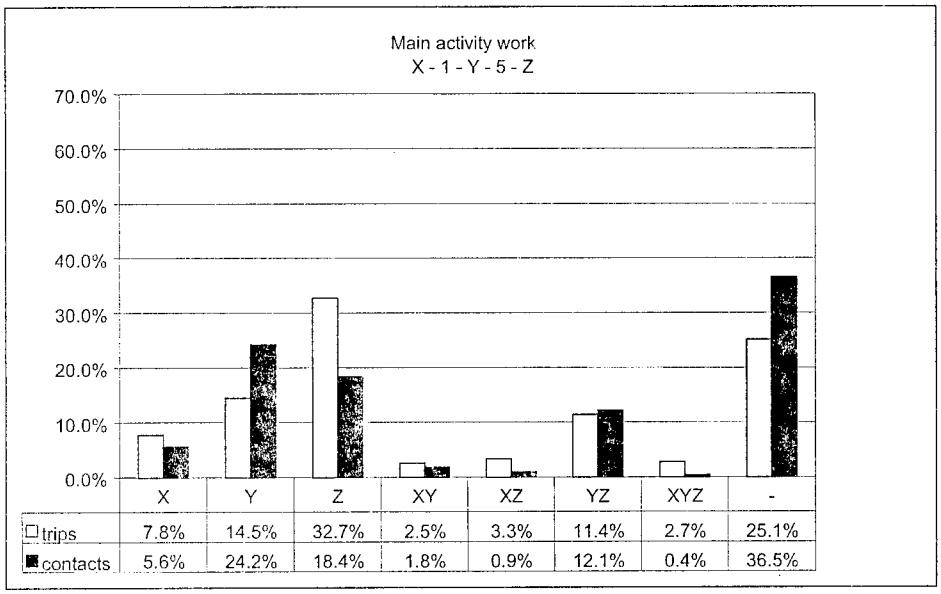


Fig. 11. Structure of trip-/contact patterns with the main activity work

In connection with the main activity 'education/training' additional trips occur mainly after returning home. This simplicity does not exist regarding the main activity 'work'. 32.7% of all trips occur after returning home from work. Further 11.4% contain additional trips between job and way home. 14.5% of all trip-/contact patterns include further trips in sequence section *Y* only. In 24.2% of all cases contacts are executed on the way to work and the way home. Again, most contacts occur in sequence section *Z*.

Due to the distribution of sequence sections the following structures for trip-/contact patterns are possible:

$$1 - 5 - Z$$
$$1 - Y - 5$$
$$1 - Y - 5 - Z$$

(For further details see also [10]).

2.3. Potential Model Developments

The detailed analysis of combined travel and telecommunication patterns indicates that a strict distinction between the substitution process (see chapter 3) and the induction process (see chapter 4: complementarity) should be made since both processes seem to be independent.

So it is to be expected that certain time-consuming, standardizable and – according to their nature – suitable activities/trips (e.g. telebanking) might be substituted in the every day life. At the same time, however, it is to be expected that an induction of physical transport is driven by new telecommunication services. The increases are to be expected fewer in normal transport processes, since the financial and time related budget restrictions are effective here due to capacity limitations of the traffic system. Increases during the weekend and holiday traffic are rather to be expected, since additional journeys can be performed neither by the generated interest via simplified information access nor by the decrease of initial travel thresholds (reduction of uncertainties concerning the selected destination by telecommunications). Within the leisure domain and the field of global business and service activities, appropriate journeys are thus generated, which would not have been possible without existence of the electronic media. And since the substitution part seems to be solvable by the introduction of transition probabilities (see chapter 3) the complementarity part needs further empirical work (see chapter 4).

3. Modelling Substitution – the Seoul Case

As part of a comprehensive research a modeling approach [8] was implemented in Seoul, for which results can be reported here. This approach was based on integrated transport/telecommunication surveys (see chapter 2.1) and interactive interviews performed in banks (telebanking) and storehouses (teleshopping). The resulting transition probabilities (reduced by the share of interviewers having had already experience, *Fig. 12*) were applied to the Transport and Telecommunication Activity Pattern by a Monte-Carlo-Simulation process. Thus, it was possible to define certain rules for the impacts of the substitution of an activity on the subsequent activities. After aggregation, the potentials to substitute trip could be quantified for the metropolitan area of Seoul (*Fig. 13*) as well as on the level of the infrastructure.

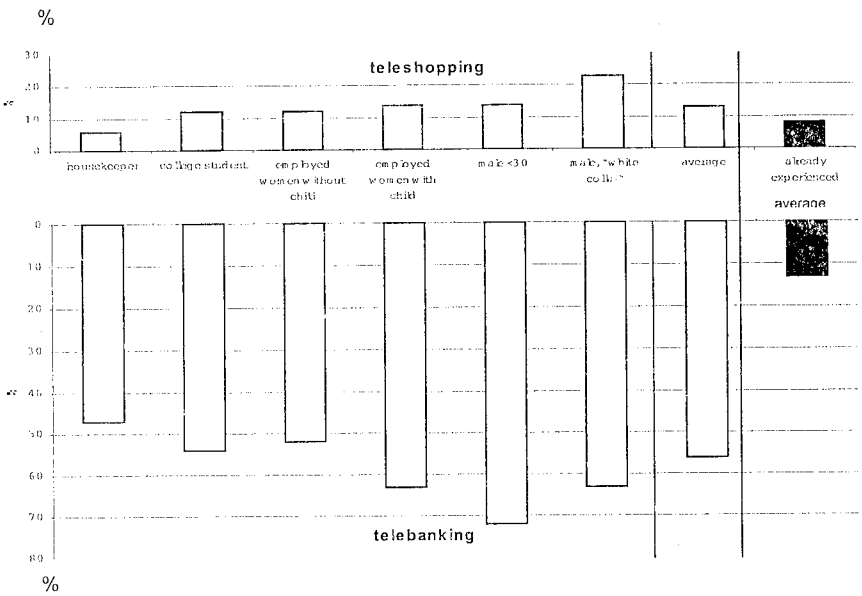


Fig. 12. Transition probabilities and experience by person groups

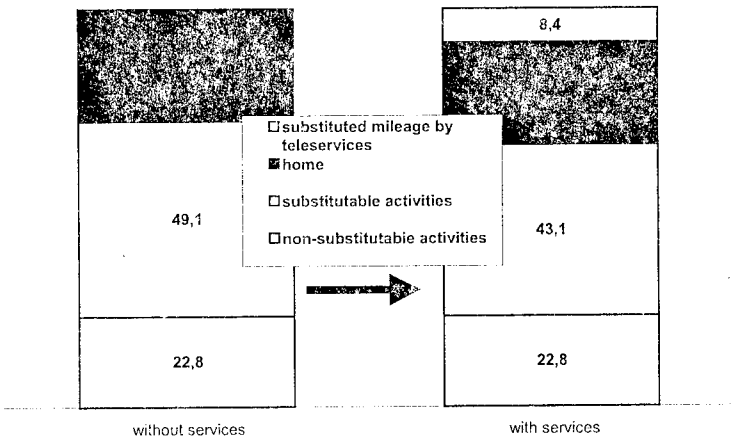


Fig. 13. Potential for substitution of mileage in Seoul

The major advantage of these simulation runs include

- the ability to run sensitivity tests related to potential measurement errors,
- the expectation to produce spatially differentiated results (infrastructure, regional development)

- the potential to include biographic factors in forecasts to reflect the different lifestyles and adaptation processes between generations (*Fig. 11*).

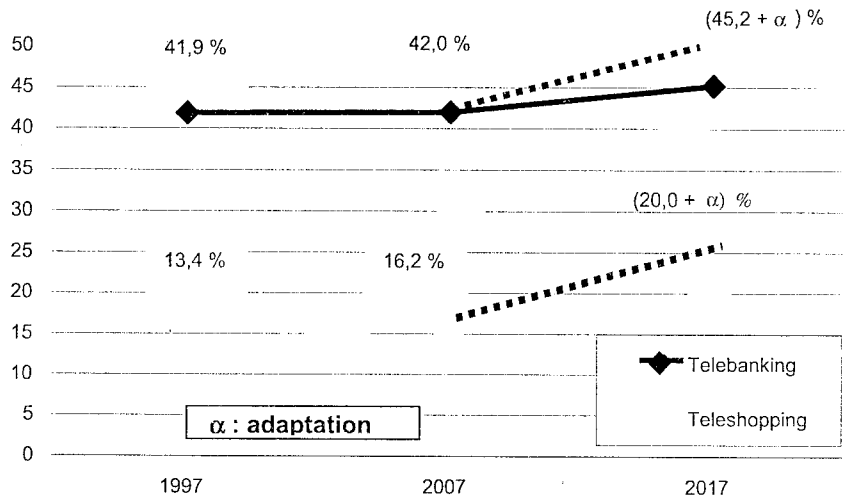


Fig. 14. The development of transition probabilities over time in Seoul

4. The Problem of Modeling Complementarity

Since decades transport researchers have worked in the field of induced demand [12], which can be seen as an analogous problem to the complementarity issue [13]. It can be described as follows: On a cross-sectional or infrastructural basis the problem seems to be solvable by the introduction of highly aggregated budgets. But on the time-axis (longitudinal view) and on an individual (biographic) basis nearly all basic questions remain without answer.

4.1. General Insight

The same situation applies to the complementarity issue. Telecommunication affects transport demands to a quite high extent, which can be quantified roughly [14]. Following our results, it can be assumed that about 40% of all contacts deal with the question of future transport – but would these trips occur without or with better means of communication to a different extent? We do not know; but we know that more empirical work may be helpful and it has to be focussed on the links which exist in combined travel and telecommunication activity patterns (see chapter 2.2) between contacts and trips. The following project shall contribute to this issue.

4.2. The EU-Spirit Project

The EU-Spirit Project intends to develop and demonstrate a customer-friendly information system for integrated travel services in Europe. Thanks to this system it will be possible, by the use of Internet as well as conventional sales and information channels, to plan and prepare a door-to-door travel between distant European regions, using all means of public transport such as main railway lines, regional trains, buses etc. The same environment offers a direct access to all other information and services which justify (e.g. event notes, tourist information) or are necessary (e.g. information on accommodations) for an attractive travel plus complete ticketing and reservation options.

Quintessential point of evaluation is an Internet-based multi-level questioning on long-distance travel behaviour before and after availability of such a Europe-wide information service. The contacting of probands takes place both 'on-trip', i.e. for example at airports, at platforms directly before a train arrives, at motorway restaurants in direct border proximity (e.g. between Germany and France), and also via Internet.

To measure the acceptance of the new information system on the basis of an interactive demonstrator in the 'sociological laboratory', the vision of the 'European information system' is obtained in form of a presentation tool (cinematic demonstration), which points out different scenarios for the future organization and the progress of long-distance travels chosen as examples, both for business travelers and for tourists. Following this demonstration a final questioning occurs on whether and to what extent this service would have affected either the specific long-distance travel via railway, plane or passenger car (depending on the location where the probands were contacted), or a trip performed recently by test persons not contacted on-trip. Additionally, those probands, specifying that such an information system would not influence their trip planning and transaction, are asked for their personal reasons for refusal. This is to permit predictions whether disuse of such a service depends either on the mode how information is provided or on a negative attitude to principle with regard to information systems, considering the previous habits.

Basic objective of the survey is the production of a quantitatively substantial database for the estimation of the long-distance travel behaviour influenced by the introduction of the information system EU-Spirit. Furthermore, to obtain insights in the process how and to which extent (novel) information contributes to the generation of transport.

5. Outlook

The outlook is completely virtual. If you would like to have a deeper insight into the structure of the interactive interviews aiming at a better understanding of the complementary contribution of telecommunication towards transport visit the

internet home page http://www.euspirit.de.

If possible, do not hesitate to participate in our latest survey on long-distance travel behaviour. And if you like to perceive whether the internet is a platform for a virtual social laboratory to support the understanding of novel teleservices, concentrate on the cinematic demonstration presented within this survey.

If you get in contact with this paper when the internet survey has already finished, ask the author (zumkeller@ifv.uni-karlsruhe.de), Dr. Stephan SCHNITTGER (sg@ifv.uni-karlsruhe.de) or Mr. Olaf EBERHARD (oe@ifv.uni-karlsruhe.de).

And please remember: Future as well as virtual reality are constituent elements of reality.

References

[1] HEINZE, W.: Zur Evolution von Verkehrssystemen, in: Perspektiven verkehrswissenschaftlicher Forschung, hrsg.: Sigurt Klatt, Dunker und Humboldt, Berlin, 1995.

[2] LÜBBE, H.: Mobilität und Kommunikation in der zivilisatorischen Evolution; in: Kommunikation und Verkehr, Fraunhofer-Forum, Tagungsband, Fraunhofer Gesellschaft, München, 1995.

[3] ZUMKELLER, D.: Transport and Telecommunication: First Comprehensive Simulation Approaches IATBR '97, *8th Meeting of the International Association for Travel Behaviour Research*, Austin, Texas, 1997.

[4] Statistisches Bundesamt, EVS, Wiesbaden 1962 – 1999.

[5] ZUMKELLER, D.: Verkehr und/oder Telekommunikation? - Konzept, Methode und Quantifizierung, Bericht der Landesarbeitsgemeinschaft Baden-Württemberg, Karlsruhe, 1998.

[6] GRÄF, P.: Dynamik der Akzeptanz telekommunikativer Dienste; in: Akademie für Raumforschung und Landesplanung, Arbeitsmaterial Nr. 251, Hannover, 1998.

[7] ZUMKELLER, D.: Communication as an Element of Overall Transport Context – an Empirical Study (Using a Modified KONTIV Method), *4th International Conference on Survey Methods in Transport*, Oxford, 1996.

[8] LEE, S.: Wechselwirkungen zwischen Verkehr und Telekommunikation in einer asiatischen Stadtumgebung, Institut für Verkehrswesen, Karlsruhe, Nr. 57, 1999.

[9] WIDLERT, S.: Survey of Communication Habits – a Short Description of Purpose and Design. Unpublished Paper by the Swedish Institute for Transport and Communications Analysis, Stockholm, 1997.

[10] SCHULMEYER, D.: Individuelles Verhalten und Aktivitätenoptimierung im Wochenverhalten, Institut für Verkehrswesen, 1994.

[11] BAUER, TH.: Kommunikationsaktivitäten und ihr Zusammenspiel mit dem Verkehrsverhalten, Institut für Verkehrswesen, Karlsruhe, 1998.

[12] POECK, M. – ZAHAVI, Y. – ZUMKELLER, D.: Traffic and Town as an Interaction Mechanism (VUSI), in: *Communications on Research Aimed at Improving Transport Conditions in Cities*, Heft 27, 1980.

[13] NILES, J. S.: Beyond Telecommuting: A New Paradigm for the Effect of Telecommunications on Travel; in: World Wide Web, 1994.

[14] MOKTHARIAN, P. L.: Now that Travel can be Virtual, will Congestion Virtually Disappear? In: World Wide Web, 1997.

PERIODICA POLYTECHNICA SER. TRANSP. ENG. VOL. 28, NO. 1–2, PP. 39–53 (2000)

USING Z SPECIFICATION FOR RAILWAY INTERLOCKING SAFETY[1]

Aleš JANOTA

Department of Information and Safety Systems
Faculty of Electrical Engineering
University of Žilina, Vel'ký diel, Žilina 010 26, Slovak Republic
Fax: +421-89-5252241
e-mail: janot@fel.utc.sk
Phone: +421-89-5133261

Received: September 29, 2000

Abstract

Formal methods involve a mature development technology that can be used to provide the highest confidence and that is used in a wide and expanding variety of environments, especially in key areas where the integrity of systems is critical or where there is a high intensity of use. Formal methods allow the logical properties of a computer system to be predicted from a mathematical model of the system by means of a logical calculation, which is a process analogous to numerical calculation. They make it possible to calculate whether certain properties are consequences of proposed requirements or whether requirements have been interpreted correctly in the derivation of a design. The objective of this paper is to discuss the possible use of formal methods in the field of requirement specification of the railway interlocking system behaviour, followed by the check of its correctness. A simple example is used to demonstrate the use of concrete formal description (Z notation). Then the principles of realisation of more complex solutions in the field of railway interlocking are indicated. Despite the specific contents of the paper concerning the railway applications the conclusions can be generalised and applied to other areas.

Keywords: formal method, Z notation, specification, interlocking.

1. Introduction

It is generally acknowledged that natural languages and similar informal notations have many disadvantages when used for writing technical descriptions – it is difficult to write specifications with the required precision, clarity and economy of expression and transform them systematically and reliably into a code or hardware. Furthermore, it is impossible to carry out formal mathematical reasoning about informally written descriptions.

Formal methods provide a framework within which people can specify, develop and verify systems in a systematic rather than ad hoc manner. They give us techniques to formally specify the functionality of a system, to verify its correctness

[1]This paper has been supported by the VEGA Grant Agency of the Slovak Republic, within the project 1/5230/98 'Theoretical Apparatus for Safety Analysis of the System with Defined Level of Safety', at the University of Žilina.

or to develop the system gradually from an abstract specification to its implementation. All these aspects are important when designing safety-critical systems.

2. Formal Methods

There are a lot of definitions for formal methods. For the purpose of this paper let us assume a definition given in (HEIMDAHL et al, 1998) and (NASA, 1998) – a method consisting of a set of techniques and tools based on mathematical modelling and formal logic that are used to specify and verify requirements and designs for computer systems and software. Such a method must provide:

- a notation with a well-defined syntax and semantics,
- some guideline and procedures for using the notation, and
- techniques for analysing specifications expressed in the notation.

Formal methods allow defects in requirements to be detected earlier than otherwise and can greatly reduce the incidence of mistakes in interpreting, formalising and implementing correct requirements. They can be used to mathematically prove that certain properties (functionality, safety etc.) hold and can be potentially applicable in all engineering disciplines, mainly for the reason that the use of formal descriptions permits machine processing of specifications. Formal methods may be classified according to different criteria: according to whether their primary purpose is descriptive or analytic (descriptive and analytic methods), according to the level of formality (with low, medium and high level of formality) and/or according to the type of the used specification language (algebraic and model-oriented).

2.1. Formal Specification and Verification

Traditionally, formal methods have been used for *formal specification and formal verification. A formal specification supports formal reasoning* which (because it can be checked by machine) can be made very reliable indeed (with reduction of human involvement) and thus it enables formal verification. In formal verification, a proof is constructed, often with mechanical support, that the specification satisfies properties of interest. Formal proofs, though, are usually very detailed. Even with machine support for their construction, formal proofs are liable to require considerable efforts to construct and check. Formal methods analysis techniques are based on deductive rather than inductive reasoning about system descriptions. These techniques include writing formal specifications, internal checking (e.g. parsing and type correctness), traceability checking, specification animation, and proof of assertions. A formal specification language typically gives a mathematical basis of a formal method. The distinction between a specification method and a specification language is fundamental. A method states what a specification must say. A language determines in detail how the concepts in a specification can be expressed. The objective of a

formal specification notation is to assist in the production of descriptions that are complete, consistent and unambiguous.

2.2. *Use, Applications and Standards*

Formal methods are used in computer science mainly for improving the quality of software or hardware or for improving confidence in the correctness of critical systems. At present they are mostly used in applications of safety critical systems, security systems, the definition of standards, hardware development, operating systems, transaction processing systems and others. For information on domains of formal method applications, see e.g. (NICHOLLS, 1992) or follow relevant discussion groups and computer science bibliographies on Internet. Due to the widening range of applications the use of formal methods is increasingly recommended by many standards (including safety ones) as a possible method of improving dependability. Highlighting the domain of railways only, the railway application standard (CENELEC, 1996) recommends formal methods (*CCS*, *CSP*, *HOL*, *LOTOS*, *OBJ*, *Temporal Logic*, *VDM* and *Z*) for development of software in a way that is based on mathematics (this includes formal design and formal coding techniques) and their use for writing specifications and for verifying the safety properties of them. Similarly, relevant standards could be identified in other application domains, too. Despite the given recommendations, there are several barriers that must be overcome of putting the formal techniques in the hands of formal developers: the notations must be more accessible and easy to use, robust tools must be produced, education must be improved and cost-effectiveness must be demonstrated. In addition, formal specification languages and their semantics are also to be standardised.

3. *Z* Notation

The *Z* notation is one of the best known formal methods which is often declared as a formal specification language and which is gaining widespread acceptance as a useful means of specifying software systems. The *Z* notation is a descriptive method with a medium level of formality. It means that the method can be used together with computer-aided tools. In the sense of a specification language it is model oriented, based upon the set theory and mathematical logic, i.e. it is based on model design which is given by description of the system state and operations over this state. The representation, structure and meaning of the formal part of specification, written in the *Z* notation, is defined in the draft standard (ISO, 1995). In the *Z* notation there are two languages: the mathematical language and the schema language. The former is used to describe different aspects of design: objects and the relationships between them. The latter is used to structure and compose descriptions: collating pieces of information, encapsulating them and naming them for re-use (WOODCOCK and DAVIES, 1996).

3.1. Aspects of Z Notation

The Z notation is based upon set theory and mathematical logic (first-order predicate calculus). Together, they make up a mathematical language that is easy to learn and apply. However, the language is only one aspect of Z. Another aspect is the way in which the mathematics can be structured. Mathematical objects and their properties can be collected together in schemas: patterns of declaration and constraint. The schema language can be used to describe system properties. A characteristic feature of Z is the use of types. Every object in the mathematical language has a unique type, represented as a maximal set in the current specification. This notion of types means that an algorithm can be written to check the type of every object in a specification; several type-checking tools exist to support the practical use of Z (e.g. $Z/EVES$, *Z Type Checker* etc.). The third aspect is the use of natural language that helps to relate mathematics to objects in the real world (partly by the judicious naming of variables, partly by additional commentary). A well-written specification should be perfectly obvious to the reader. The fourth aspect is refinement. We may develop a system by constructing a model of a design and then refine the description by constructing another model which respects the design decisions made, and yet is closer to implementation. Where appropriate, this process of refinement can be continued until an executable code is produced. Thanks to given aspects the Z notation is a mathematical language with a powerful structuring mechanism.

3.2. Trivial Application Example

For the sake of better understanding, a trivial example of Z specification is presented, based on (JANOTA, 2000), to illustrate a typical process of formal specification design:

- Giving an informal problem statement of a system: this step consists in informal specifying the system requirements, using informal means of expression (natural language, diagrams, graphs, tables etc.).
- Making Z style of presentation: Z specification consists of many small sections, each beginning with informal descriptions followed by formal texts. Natural order for easy understanding is from top down, from general to detail, from common to exception and avoids forward referencing. If necessary, larger entities can be built. Z adopts the state machine view of systems (*Fig. 1*).
- Writing Z specification using a Z-based tool: in this step the Z specification is transformed into the required computer-based format.
- Performing tool-supported activities: a variety of activities (pretty-printing, editing and browsing, type-checking, consistency checking, proof-obligation generation, theorem proving, code generation) can be available, based on the tool to be used.

3.2.1. *Informal Problem Statement*

The task consists in developing a system (hardware or software) whose behaviour makes substitution of electro-mechanical relay possible and whose specification is the task. There are different ways of how to describe a system behaviour informally – using verbal descriptions, specifying a voltage dependency that represents the hysteresis characteristic of relay operation, writing mathematical equations, making logical decision tables etc.

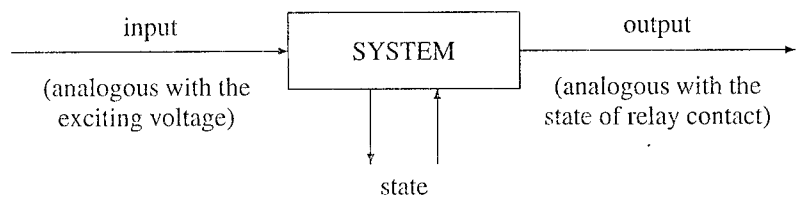


Fig. 1. State machine view to the system substituting an electro-mechanical relay

Behaviour of the system in *Fig. 1* can be described verbally:

a) If the (internal) state of the system is analogous with the relay drop-away, the imaginary relay pick-up will occur if the input voltage reaches the value equivalent to or higher than the nominal (just-operate) pick-up voltage (U_{pn});

b) If the internal state of the system is analogous with the relay pick-up, the imaginary drop-away will occur if the input voltage falls down to the value equivalent to or lower than the nominal drop-away voltage (U_{on}).

For the particular relay type, the value can be specified e.g. $U_{pn} = 15$ V and $U_{on} = 10$ V with the input voltage range 0 to 48 V.

3.2.2. *Formal Z Specification*

In this section a way of Z specification design is presented with brief explanation of particular parts. More details about Z syntax and semantics are available e.g. in (WOODCOCK and DAVIES, 1996) or (ISO, 1995).

$$States ::= pick_up/drop_away \tag{1}$$

Eq. (1) is called a free-type definition, which introduces a set of all possible states of the system containing two distinct elements. The order in which these elements are introduced is unimportant.

$$\begin{array}{|l} voltage,\ U_{pn},\ U_{on} : \mathbb{Z} \\ \hline U_{pn} = 15 \\ U_{on} = 10 \end{array} \tag{2}$$

Axiomatic definition (2) introduces new symbols in its declarative (upper) part as elements of the set $\mathbb{Z}$. In the lower (predicate) part there are constraints – in our case the nominal voltage values. Such a definition is said to be axiomatic, as the constraint is assumed to hold whenever the symbol is used: it is an axiom for the object.

$$Response ::= close/open \tag{3}$$

Eq. (3) gives a free-type definition of the system response. Let's suppose a response analogous with the behaviour of the relay contact system (contacts *closed* or *open*).

$$\begin{array}{|l} \hline \textit{System} \\ state: States \\ voltage: \mathbb{Z} \\ \hline voltage \in 0..48 \\ \hline \end{array} \tag{4}$$

The system state is described by the so-called vertical definition of the state schema (4). The schema consists of the following parts: the name of schema (*System*), declaration (expressed as *name: type*) and constraints. The state schema defines all the allowable or valid system states (sometimes called cases). The system state is allowed if all the conditions in the axiomatic part are satisfied. As an example let's assume a particular situation where the system is in the state close and the *voltage* is 30 V.

$$\begin{array}{|l} \hline \textit{InitSystem} \\ System \\ \hline state = drop_away \\ voltage\ =\ 0 \\ \hline \end{array} \tag{5}$$

The schema (5) says how the system should be initiated. The internal state is defined as drop away together with the zero input voltage. After initiation we specify the normal situations first. Every operation is specified using an *operational schema*.

PickUp

$\Delta System$
$voltage? : \mathbb{Z}$
$response! : Response$

$state = drop_away$
$voltage? \geq U_{pn}$
$state' = pick_up$
$response! = close$

(6)

Operation (6) is analogous with the relay pick-up. According to Z conventions, $\Delta System$ indicates that the operation *PickUp* may modify the state. The name *state* refers to the state before the operation (called the *pre-state*), the dashed name *state'* refers to the state after the operation (called the *post-state*). Suffixes '?' and '!' indicate an input name and an output name, respectively. The schema (7) describes the situation analogous with the relay holding ($\Xi System$ indicates that the system state is unchanged) where its output remains *close*.

Holding

$\Xi System$
$voltage? : \mathbb{Z}$
$response! : Response$

$state = pick_up$
$voltage? > U_{on}$
$state' = state$
$response! = close$

(7)

$$
\begin{array}{|l}
\hline
\textit{DropAway} \\
\Delta System \\
voltage? : \mathbb{Z} \\
response! : Response \\
\hline
state = pick_up \\
voltage? \leq U_{on} \\
state' = drop_away \\
response! = open \\
\hline
\end{array}
\tag{8}
$$

$$
\begin{array}{|l}
\hline
\textit{NoPickUp} \\
\Xi System \\
voltage? : \mathbb{Z} \\
response! : Response \\
\hline
state = drop_away \\
voltage? < U_{pn} \\
state' = state \\
response! = open \\
\hline
\end{array}
\tag{9}
$$

Schemas (8) and (9) describe other situations. Finally, there are so-called *horizontal definitions* (10) and (11). Their names correspond to states analogous with states of the relay contact system.

$$Close \triangleq PickUp \vee Holding \tag{10}$$

$$Open \triangleq NoPickUp \vee DropAway \tag{11}$$

Let's consider the specification definitive though it could be developed and refined.

3.2.3. Machine Processing of the Specification

The computer-based format of the specification is given by requirements of the tool used to check syntax correctness (alphabet of symbols and grammar rules

which define well-formed formulations) and semantics (mathematical proof of the contents correctness). A particular semantic construction is a model of a particular Z specification (or predicate) if the relationship between terms in the specification also holds in the model (if it makes the predicate true). A predicate is logically valid if it holds in all models. It is logically valid relative to a specification if it holds in all models of that specification. Logic for Z is a collection of axioms and rules of inference written using the Z notation. These rules and axioms may be used inductively to define which predicates are logical consequences of others. Those, which are the logical consequence of the empty set of predicates (or, equivalently, the logical consequence of only the axioms of the logic), are known as theorems. A logic for Z may be proven sound in some model by showing that each of its theorems is logically valid. It is traditional to ask whether a logic is also complete, i. e. whether every logically valid predicate is also a theorem. More details on distinguishing between the different philosophies of proof in Z are available in (MARTIN, 1999).

All the specifications presented in this paper were transformed to the format required by the software package *Z/EVES ver. 2.1* by *ORA Canada*, running under *MS Windows 98*. The tool uses its own file format (*.*zev*, *.*zcs*) or can import the files in LaTeX2e format (*.*tex*). *Z/EVES* has two parts: a server that checks paragraphs and executes proof commands, and a GUI that manages specifications, sends commands to the server, and displays results. An example display of the specification window is given in *Fig. 2* and relates to the contents of Section 4.

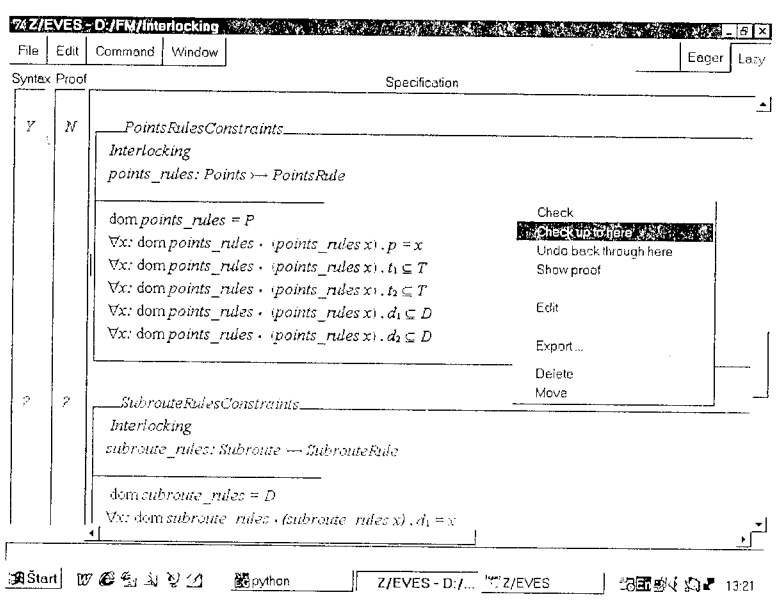


Fig. 2. Example of a *Z/EVES* specification window

The status of each paragraph is shown in two columns to the left of the paragraph. The leftmost column shows one of three symbols:

- '?' indicates that the paragraph has not been checked,
- 'Y' indicates that the paragraph has been checked and has no syntax or type errors,
- 'N' indicates that the paragraph has been checked and has errors.

The next column shows the proof status for the paragraph, using one of three symbols:

- '?' indicates that the paragraph has not been successfully checked (so proof status cannot be determined),
- 'N' indicates that the paragraph has unproved goals,
- 'Y' indicates that the paragraph has no unproved goals.

4. The Use of Formal Specification in Railway Interlocking

Modern computer interlocking systems typically consist of hardware and software where the software generally consists of:

- Generic programmes, common for different installations,
- Geographic programmes, considering local data.

Correctness of requirement specification must be assured in order to achieve functional safety of the interlocking system. Due to its high complexity it is apparent that verification of functional dependencies is an extremely complex task. Since the internal state has extremely many possible configurations there is an effort to perform the automatic analysis of the safety features of programmes with local (geographic) data. This task can be solved with the help of formal specification and verification methods.

4.1. Starting Points of Formal Analysis

Following the methodology introduced in the previous section the first step consists in informal problem definition. Any railway authority defines own regulations and standards concerning the rules for railway vehicle movement. Under conditions of ŽSR (Slovak National Railways) the relevant requirements are given in the form of government and technical standards (mainly of the series TNŽ 34 26xx) and other relevant regulations. Developing a concrete application these rules are applied to the given railway topology. This results in the table form containing all the possible operational situations. From our viewpoint this can be considered to be an informal requirements specification for the interlocking under specific local conditions. Not all railways use such table forms, some substitute them with alternative descriptions

of possible situations through rules or logical equations for route settings (with final document called User Case). However, the final specification is a product of manual activity of the staff and is subjected to strict validation and certification process. This process may be automated with the use of formal methods.

4.2. State of an Interlocking System

The state of an interlocking system is generally defined by combination of particular states of its components. These include mainly the following (physical) objects:

- Track circuits, dividing the track into sections and detecting their states (occupied or clear),
- Points, steering trains across junctions and finding themselves in one of defined positions (controlled plus, controlled minus or undefined),
- Signals, allowing or refusing the entry of the railway vehicle onto particular sections of the track and situated in advance of the section which they control.

Apart from the previous (physical) objects there are also so-called logical objects:

- Routes as sections of track between two signals, which proceed from an entry signal to an exit signal (the route set or unset),
- Sub-routes as subsections of routes that are associated with specific track circuits. The concept of the route as a set of sub-routes is typical for several computer interlockings. The sub-route can be locked or free.

The process of defining the local dependencies (based on geographic railway topology) most often concentrates on the following invariant conditions (SIMPSON, 1998):

- For every track circuit, no more than one of the sub-routes passing over it should be locked for a route at any time,
- If a sub-route over a track section containing points is locked for a route, then the points are correctly aligned with that sub-route,
- If a route is set, then all of its sub-routes are locked,
- If a track circuit containing points is occupied, than the points are locked,
- If a sub-route is locked for a route, then all sub-routes ahead of it on that route are also locked.

If any of these invariants are not satisfied, then there is an error in the geographic data. The task defined in the way above enables to automate the process of relevant data model design (geographic database) and verify its safety.

4.3. Concept of Z Representation

Within the first step the sets of particular components are defined. In the Z notation, there are several ways of defining an object – declaration, abbreviation, generic abbreviation, axiomatic definition and generic definition (WOODCOCK and DAVIES, 1996). We use the first way, declaration, to define new basic types, representing components of the interlocking:

$$[TCircuit, Points, Signal, Route, Subroute] \tag{12}$$

The schema *Components* describes the sets of track circuits, points, signals, routes and sub-routes with which a particular interlocking is concerned.

Components

$$\begin{aligned} &T : \mathbb{P}\, TCircuit \\ &P : \mathbb{P}\, Points \\ &S : \mathbb{P}\, Signal \\ &R : \mathbb{P}\, Route \\ &D : \mathbb{P}\, Subroute \end{aligned} \tag{13}$$

Relationship among different components of an interlocking is specified by the schema (14).

Structure

$$\begin{aligned} &locks: Route \rightarrowtail\!\!\!\!\!+ \mathbb{P}\, Points \\ &sroute: TCircuit \rightarrowtail\!\!\!\!\!+ \mathbb{P}\, Subroute \\ &berth: Points \rightarrowtail\!\!\!\!\!+ \mathbb{P}\, TCircuit \\ &tcircuit: Subroute \rightarrowtail\!\!\!\!\!+ TCircuit \\ &sroutes: Route \rightarrowtail\!\!\!\!\!+ \mathrm{seq}\, Subroute \\ &plus: Points \rightarrowtail\!\!\!\!\!+ \mathbb{P}\, Subroute \\ &minus: Points \rightarrowtail\!\!\!\!\!+ \mathbb{P}\, Subroute \end{aligned} \tag{14}$$

Components and *Structure* together describe the layout of an interlocking. Points can be in one of two positions: controlled plus (cp) or controlled minus (cn). Therefore, we introduce the free type *points_position* (16). A point can be controlled free to go plus (cfp) or controlled free to go minus (cfm) or both at any time. Thus, the free type *points_state* is introduced by (17).

The free type *tcircuit_state*, *route_state* and *sroute_state* are introduced similarly. Track sections, with train detection devices, inform the interlocking if a

section is occupied (o) or clear (c), (18). A route may be in one of two states (19): set (s) or unset (us). Sub-routes are subsections of routes, which are associated with specific track circuits. They may be in one of two states (20): locked (l) or free (f).

Interlocking

Components

Structure

$$\begin{aligned}
&\mathrm{dom}\ locks = R \wedge \bigcup(\mathrm{ran}\ locks) \subseteq P \\
&\mathrm{dom}\ sroute = T \wedge \bigcup(\mathrm{ran}\ sroute) \subseteq D \\
&\mathrm{dom}\ berth = P \wedge \bigcup(\mathrm{ran}\ berth) \subseteq T \\
&\mathrm{dom}\ tcircuit = D \wedge \mathrm{ran}\ tcircuit = T \\
&\mathrm{dom}\ sroutes = R \wedge \mathrm{ran}\ (\bigcup(\mathrm{ran}\ sroutes)) \subseteq D \\
&\mathrm{dom}\ plus = P \wedge \bigcup(\mathrm{ran}\ plus) \subseteq D \\
&\mathrm{dom}\ minus = P \wedge \bigcup(\mathrm{ran}\ minus) \subseteq D
\end{aligned} \tag{15}$$

$$points_position ::= cp \mid cm \tag{16}$$
$$points_state ::= cfp \mid cfm \tag{17}$$
$$tcircuit_state ::= c \mid o \tag{18}$$
$$route_state ::= s \mid us \tag{19}$$
$$sroute_state ::= f \mid l \tag{20}$$

To formalise the structure of a route setting rule, a number of entities must be concerned: the route which is to be set (r), a set of conditions on points which must be satisfied before the route can be set (p_1), a set of point movements form part of the setting process for the route (p_2), a set of sub-routes which must be free before the route can be set (d_1), and a set of sub-routes which must be locked when the route is set (d_2). Then the route setting rule can be specified (21).

RouteRule

$$\begin{aligned}
&r : Route \\
&p_1 : \mathbb{P}\ (Points \times points_state) \\
&p_2 : \mathbb{P}\ (Points \times points_position) \\
&d_1 : \mathbb{P}\ Subroute \\
&d_2 : \mathbb{P}\ Subroute
\end{aligned} \tag{21}$$

The schema (21) describes a sub-route release rule. This involves a set of track circuits which are required to be clear, a (possibly empty) set of sub-routes which are required to be free, and a (possibly empty) set of routes which are required to be unset.

Finally, we can also represent the rule for points (23) where p is the point with which we are concerned, t_1 is the set of track circuits associated with the free to go plus condition, t_2 is the set of track circuits associated with the free to minus position, d_1 is the set of sub-routes associated with the free to go plus condition, and d_2 is the set of sub-routes associated with the free to go minus condition.

SubrouteRule

$d_1 : Subroute$
$t : \mathbb{P}\, TCircuit$
$r : \mathbb{P}\, Route$
$d_2 : \mathbb{P}\, Subroute$

(22)

PointsRule

$p : Points$
$t_1 : \mathbb{P}\, TCircuit$
$t_2 : \mathbb{P}\, TCircuit$
$d_1 : \mathbb{P}\, Subroute$
$d_2 : \mathbb{P}\, Subroute$

(23)

With restriction to the three types of rules with which we are concerned, the overall structure of a geographic database may be represented by the schema (24). This contains three schemas, which specify detailed conditions (constraints) for each rule and are not specified here.

Database

RouteRulesConstraints
SubrouteRulesConstraints
PointsRulesConstraints

(24)

5. Conclusions

There are different reasons for applying formal techniques: an effort to improve quality of the whole development process, to improve integrity, reliability or other

characteristics of the system, to reduce specification errors, to improve requirements definition and documentation, to provide a firmer foundation for maintenance and enhancement and a more rational basis for choosing test data, to explore properties of a design architecture, to be as certain as possible that the design and implementation are error-free, to meet particular customer or standard requirements. The model-based approach results in a model of the system that acts as a specification and is constructed using well-understood mathematical entities. Setting out pre and post conditions for the function can specify functions. However, this approach does not scale up to large or medium sized systems. Thus, the Z notation is not intended for the description of non-functional properties, such as usability, performance, size and reliability. Neither is it intended for the description of timed or concurrent behaviour. This brings certain limitations when using it for formal specification of railway interlocking systems. However, there are other formal methods that may be used in combination with Z to relate state and state-change information to complementary aspects of design. For such purpose there are often used *Communicating Sequential Processes* (CSP). This formalism is able to describe concurrent behaviour of more objects or processes through interaction with their environment. More details about the way of modelling a geographic database using this formalism and checking it by $FDR2$ (*Failures-Divergences Refinement2*) can be found e.g. in (SIMPSON, 1998). Despite all discussed limitations, the functional view of using Z is widely accepted, and combined with the behavioural view supported by other formal techniques rather than used independently.

References

[1] CENELEC European Committee for Electro-technical Standardisation, Railway Applications: Software for Railway Control and Protection Systems. Draft prEN50128, August 1996.

[2] HEIMDAHL, Mats P. E. – HEITMEYER, Constance L., Formal Methods for Developing High Assurance Computer Systems: Working Group Report, *Proc. 2nd IEEE Workshop on Industrial-strength Formal Techniques (WIFT'98)*, Boca Raton, FL. October 19, 1998.

[3] ISO SC22 Working Group 19, Z Notation. Technical report, ISO/IEC JTCS1, SC22 N1970. ISO CD 13568, Committee Draft of the proposed Z standard, 1995.

[4] JANOTA, A., Formal Specification of a Railway Interlocking System. *Studies of the Faculty of Electrical Engineering*, University of Žilina, **26** (2000), pp. 13–25. (In Slovak).

[5] MARTIN, A., Relating Z and First-Order Logic. In: J. Wing, J. Woodcock, J. Davis (Eds.): *Proc. FM'99*, **II** (1999), LNCS 1709, pp. 1266–1280.

[6] NASA/TP-98-208193, NASA Formal Methods Specification and Verification Guidebook for Software and Computer Systems, Vol. I: Planning and Technology Insertion. 88 p, 1998.

[7] NICHOLLS, J. E., Domains of Applications for Formal Methods. *Proc. of Z User Workshop*, University of York, Workshops in Computing, Springer-Verlag, 1992.

[8] SIMPSON, A. C., Model Checking for Interlocking Safety. *Proc. FMERail Workshop*, Vol. 2 of 2nd FMERail Seminar, Parks Road, Oxford OX1 3QD, England, October 1998, ESSI Project 26538.

[9] WOODCOCK, J. C. P. – DAVIES, J., Using Z: Specification, Refinement and Proof. Prentice Hall International Series in Computer Science, 1996.

PERIODICA POLYTECHNICA SER. TRANSP. ENG. VOL. 28, NO. 1–2, PP. 55–64 (2000)

ADOPTION OF EUROPEAN UNION'S LAW IN HUNGARY, IN THE AREA OF PRODUCT ADEQUACY

Jenő FRANK and András GYÖRGY

Department of Building Machines,
Materials Handling Machines and Industrial Logistics
Faculty of Transportation Engineering
Budapest University of Technology and Economics
H–1521 Budapest, Hungary
Tel: +36 1 463 2235; Fax: +36 1 463 3059
e-mail: frank-jeno@eagt.bme.hu

Received: Nov. 1, 2000, Revised: Dec. 20, 2000

Abstract

The EU measures to introduce the internal market came into force on December 31st, 1992. The Council of the European Community produced a Directive, which describes harmonization of rules issued by the Member Countries concerning machines. The European Union is an important economic partner of Hungary. We analyzed the situation concerning machines before the Hungarian connection to European Community.

Keywords: harmonization, standardization, accreditation, safety, product warranty, lift.

1. Technical Regulatory System within the European Union

Current economic trends are showing that both Europe and the world is being transformed into a huge, more or less open market economy. Within that scope, single countries shall play a significant or an insignificant role in proportion of their respective economic potentials. No country may refrain itself from the existence of market mechanisms. This mechanism is functioning on the basis of the cooperation of 'compatible' economies of individual countries, as an organizing principle. One condition of this compatibility is the common or at least harmonized system of legal regulations, norms and standards concerning the economy, the safe functioning and proper quality of products. In similar cases, even products shall be used in common with other products or parts originating from different countries, in that case, compatibility shall be interpreted in its even narrower sense, too.

The EU measures to introduce the internal market came into force on December 31st, 1992. Internal market means a territory without internal limits, where free movement of goods, persons, services and capital is guaranteed. The European Union has become the biggest consumer market in the world in terms of number of people (if China is not considered). At the same time it has become – behind the United States – the second producer of goods. Technical harmonization within the EU shall achieve its purposes, when a product manufactured according to the

standards of a country fulfils the prescriptions of the norms of any other country without any modification and vice versa.

Realization of the aims within the unified European market has involved the necessity to form a quality policy at EU level. Breakdown of outside customs (technical) hindrances forced the Community to develop its own system of technical regulations. The model is based on legal harmonization and mutual acceptance of national standards specifications. Consequently, EU technical regulations are subdivided into two areas: legally regulated and market regulated (spontaneous) spheres.

1.1. Legally Regulated (Harmonized) Sphere

This sphere expresses requirements in harmonized technical legal regulations. EU legal practice is hereby confined to fixing the essential requirements which should be respected by products and services in order to ensure the protection of health, safety, environment as well as the customer itself. In these fields technical solutions cannot be selected by the manufacturers concerned at their will, the frameworks shall be ensured by harmonized legal regulations, called directives.

The Machine Directives [1] has been in force since 1995 and enacted by each EU member country as a national law. Harmonization of legal rules involves more and more fields, the list of products concerned cannot be considered to be final and definitive. One product may be included in more directives as well: lifts e.g. by the directives for machines, low tension electric devices and electromagnetic compatibility. The basic requirements expressed by directives may be completed by special requirements as well.

1.2. Market Regulated Sphere

In the market regulated sphere, the requirements concerning goods and services are regulated by specification documents (norms, contracts) voluntarily accepted by the participants of economic life instead of legal rules. Within the scope of mutual acceptance, member countries should permit the free access to the market of any product manufactured and commercialized under lawful conditions in any other member country.

Within the EU, in order to ensure the function of a unified internal market, common standard minimum conditions are fixed for each requirement to be asserted, beyond which no other conditions can be claimed by the authorities of any country.

2. EU Rules Concerning Machines

The scope of the Directive of the Council of the European Community [1] describes harmonization of rules issued by the Member Countries concerning machines. For some sorts of machines, harmonized EU Directives (and ISO and EN Standards as well) concerning prescriptions for designing and manufacturing of them have been issued. Under the regulation, a machine can be distributed if its functioning meets the necessary safety criteria. Safety regulations should be harmonized in order to provide the free movement of machines without a drop in safety levels in the member states.

Social expenses related to the great number of accidents during the functioning of the machines could be reduced by a designing and manufacturing process respecting safety regulations in the first place, as well as by appropriate maintenance.

The Directive includes some partial safety equipment sold separately as well. Some machines are not included into the scope of the Directive, e.g. those operating exclusively by direct human force, ropeways, paternoster lifts, theatre traps, building site lifts, etc. Building and transporting machines not listed here shall be included into the Directive.

Annex 1 of the Directive contains the basic health protection and safety requirements concerning the designing and manufacturing of machinery and safety parts, in particular, passenger lifts and transporting equipment.

Strict quality control has been introduced concerning particularly dangerous machines. All machines bearing the CE sign may move and be put into operation freely in the EU. Member Countries shall not prohibit, restrict or hamper in their territory such machines and safety parts to be commercialized and put into operation, which correspond to requirements.

If a machine out of standard bears a 'E' sign or an out of standard safety part is accompanied with a declaration of EU-conformity, the member country concerned shall to initiate an appropriate process against the person having issued the sign or the declaration and inform the EU committee and the other member countries thereof.

3. Functioning of Certification Organs

In the legally regulated sphere, certification shall be carried out by the so called notified bodies.

These are appointed by member countries - from technically competent organs under their legal authority – at their own responsibility, then notified to the EU Committee and to other member countries. Notified organs shall be registered at EU level, the list is published in the Official Journals of the Community and kept up-to-date.

In the scope of harmonizing legal regulations concerning machines, the

Directive of the Council of the European Community [1] deals with the problem of minimum criteria for appointing certification organs, which shall be respected by member countries:

- the certifying organization shall not be a designer, manufacturer, supplier of the machine to be tested,
- highest level of technical competence,
- technical tasks involved to tests,
- impartiality of persons carrying out tests,
- the certifying organization shall take out a third party insurance, except if responsibility is taken over by the state,
- professional secrecy.

Main rules of evaluating conformity: The manufacturer shall issue an EU Adequacy Declaration to each machine or safety part manufactured and place the CE sign onto the machine. Elevators are classified here in the group of machines qualified as particularly risks with e.g. sawing machines, presses, some machine tools for underground works.

Even safety part equipment shall be subject to the same certification procedure applied to machines. If neither the manufacturer, nor its representative within the EC fulfils its commitment concerning the certification process, responsibility is taken over by all those persons that distribute the machines or the safety equipment within the EC.

4. The Agreement of Association

The European Union is an important economic partner of Hungary. Companies established in Hungary with foreign capital have been founded merely by companies established within the EU. Relations are built at the level of interstate agreements, political and economic measures and company actions.

The Parliament of the Republic of Hungary has promulgated by its Law I of 1994 the European Agreement about establishing an Association with the European Community. This Association aims among others at

- a free trade zone to be introduced stepwise – within 10 years – between Hungary and the Community;
- an improvement to be achieved between the Parties in the realization of other economic liberties the Community is based on;
- economic, financial and cultural cooperation to be promoted at the widest possible level.

The Agreement indicating a schedule of stepwise reduction of customs duties in its Chapter dealing with free movement of goods, while setting up the framework of economic cooperation – among others – aims at promotion of industrial cooperation, development and protection of investments in the fields of science and technology,

education and training, environmental economy and traffic. Here is the cooperation in the fields of industrial standards and quality assurance dealt with, where the diminishing of existing differences is aimed at, for this end·

- promoting the application of community's technical regulations, European standards and quality control processes;
- supporting Hungary's participation in the European programs concerning measurements and testing;
- supporting Hungary's participation in the specific organizations (CEN, CENELEC);
- promoting the exchange of technical and methodical data among the international parties in the field of quality control.

In justified cases, the Community shall provide technical assistance to Hungary.

5. Certification of Adequacy in Hungary

5.1. The Safety Code

The Labor Safety Code [3] being the most important Hungarian legal regulation dealing with safety technology regulates this very far-reaching field in harmony with the Law of the European Union.

Any tools can be sold, imported, put into operation or used only if the requirements of working conditions are safe and not endangering health, which should be certified in the framework of general quality certification by the manufacturer, in the case of imported goods, the importer, in lack of importer, the user.

In the enacting clauses of the Minister of Labor [4], sale, implementation and use of the listed tools, as well as of any personal safety device are subject to a valid certificate of adequacy. Safety adequacy shall be tested as set forth by the Law on Quality and by the Decree of the Minister of Labor.

In case the manufacturing or implementing particular devices or tools is subject to a license issued by the competent authorities (such as elevators), the license is equal to a certificate of conformity.

In the above Decree of the Minister of Labor, the list of tools and devices, which shall be operated exclusively with a certificate of adequacy, includes among others: cranes, lifts, ropeways, mobile or other personal lifts, etc. Safety adequacy of tools and devices listed, when being of national origin, shall be certified by a second level certification as per a separate legal regulation about quality systems (not issued yet).

Remark: The Decree of the Minister of Labor not issued yet regulates the process of granting certificates of adequacy to tools, devices and establishments not subject to a license, and the (national or EU-) notified organizations authorized to issue such a certificate in Hungary.

In the case of imported tools listed in the Decree, safety adequacy shall be certified by an equivalent certificate of quality in Hungarian in the form as accepted by the Hungarian Standards Institute, or by its legal successor, the Hungarian Standards Corporation or a Certificate of Adequacy of an accredited Test Laboratory shall be produced.

Safety shall be periodically supervised (in the following referred to as 'Supervision') for tools and devices included in the above list, whose operation is subject to a certificate of adequacy, furthermore for any machine, whose manufacture or putting into operation is subject to a license issued by competent authorities, or the Supervision of which is foreseen by Law, by Standards Specification or by the documentation of operation.

The EU Directive on machines [1] is now valid in Hungarian law, too: in form of Decree of the Minister of Industry, Tourism and Trade [13].

5.2. Process before Authorities in Concern of Lifts

The Decree of the Government [12] about establishing, operating and testing lifts is dealing with lifts and charge elevators in any kinds of buildings, furthermore with paternosters for the transport of persons, escalators and travolators.

The Decree does not cover mine lifts, special external transporting equipment, cable-cars, ski lifts, ropeways, theater trap door and other elevator equipment, lifts in explosion dangerous area, spindle lifts, braking gear equipment, furthermore such hoisting and lifting equipment, where the car leaves guiding rails to cross to another railway, finally hoisting, transporting and loading equipment for industrial technologies or machine tools.

The competent authorities shall be requested to issue a license to build, to move or to modify a lift equipment if principal technical data shall be altered. Before submitting an application thereof, an expertise from the Quality Inspection Innovation Inc. of Building Industry (ÉMI Inc.) shall be enclosed stating that the lift is suitable for use. Both the contents of lift documentation and lift logbook is fixed in the Decree.

First instance licenses are issued by the notified bodies (this time there are 2 notified bodies in Hungary). Lift operators are liable to command yearly supervision works, to have defects and faults detected by the notified bodies, to report offsetting or re-implementation of lifts to ÉMI Inc.

The names and addresses of lift experts of ÉMI Inc. are registered with first instance building authorities. There is a national register of all lifts kept by ÉMI Inc.

On the basis of the above it can be stated that ÉMI Inc. is entitled in the territory of Hungary to issue certificates of adequacy in the field of lifts.

The Hungarian standards are equivalent with the Euronorms (EN 81 and EN 115), too.

5.3. Certification of Quality

Certification means that an independent third party acknowledges that a product, service or person complies with a standard or with any other prescription.

Certification may concern products, the quality system of the company, but in some cases even company personnel shall be certified, e.g. in the case of welders.

In the European Union in the fields governed by legal prescriptions it is compulsory to carry out adequacy tests for being introduced to market.

In the fields governed by market trends there is no legal prescription for being certified. Manufacturing and service companies shall – with regard to expectations both from market and customers – decide upon their own being certified and upon their standard level. Certification means here an adequacy to standards, to agreements as accepted voluntarily.

Building up quality assurance systems and their certification according to a standard of ISO 9000 series shall be a preliminary condition of getting to market to an even wider extent. An overwhelming part of certified systems is to be found in Europe.

Companies intending to export to the EU market shall be certified according to a standard of ISO 9000. This requirement has been introduced – to replace customs duties abolished – as a 'fine tuning' tool for market protection within the EU.

However, the importance of a certificate according to ISO 9000 on the market shall not be overestimated; it shall be considered as a necessary but not satisfactory condition of competitiveness. The certificate acknowledges that the company has carefully considered and documented its activity in relation to quality, by this becoming transparent and controllable, and if some faults are detected in the field of quality, corrective measures are taken. Neither high technical level, nor product quality is hereby guaranteed!

6. Methods of Product Evaluation to Establish Conformity

The Decision of the Council of the European Communities [5] deals with different methods of adequacy evaluating procedures. This Decision has fixed the activity during adequacy tests of both manufacturers and evaluating organizations in module elements.

- Modules are marked from A to H and concern both manufacturing and design phases of products.
- Directives list modules to be selected for a given product or product group and tend to offer the widest possible selection for the manufacturer to comply with requirements. Modular approach is a manufacturer-friendly solution never forcing the manufacturer to choose a particular preferred solution to have the adequacy of its product or service certified.

- Directives contain criteria for the manufacturer to consider when selecting the most appropriate modules from the list given in the Directives for the manufacturing environment of its product. Among them, the risk involved in the product is of basic importance.

7. Accredited Organizations

Testing is a sort of service. For this reason, it is extremely important that the customer, the sales partner have confidence in the certifying organization. Accreditation helps to reinforce confidence and mutual acceptance of certificates of conformity.

Accreditation is an official acknowledgment of the fact that an organization or an institution is prepared to exert certain activities (tests, certification, inspection) under certain conditions (legal prescriptions, standards specifications, etc.).

Accreditation can be applied both in fields governed by legal prescriptions or by market conditions. In former case, ministries decide whether accreditation is a primary condition of being appointed to carry out authority tests or to certify. According to EU-recommendation, appointment shall be based on accreditation. Therefore the trend is that countries are building up their national accreditation systems and they take them into consideration in the work of granting licenses by authorities, too. In the sphere governed by legal rules, exclusively state appointed organizations can be accredited, in the fields governed by market tendencies any organization may apply for being accredited.

National accreditation is a procedure carried out by an official, authorized organ of a state (in Hungary, the National Accreditation Body [9]) deciding upon suitability to certify, to test. In the EU, there is no unified accreditation system. In the most developed countries (e.g. Great Britain, Germany, France), accreditation has got a definite system, while in other countries (e.g. Greece, Belgium) there are no national accreditation systems. No international accreditation top organization is planned within the EU. As a goal, mutual acceptance of certificates issued by accredited certifying organizations was expressed.

8. National Quality Certification System

The National Standardization Law [8] sets forth as a task of the Hungarian Standards Corporation (MSZT) to elaborate the rules for using national standards signs and the conditions of their granting. Firms may apply voluntarily – in the hope it brings advantages on market – for the adequacy of their products to definite standards certified by MSZT. About products entitled to use the sign 'Tested for ade- quacy with MSZ ...' an up-to-date information can be obtained from MSZ with indication of manufacturer or dealer, furthermore, the Official Journal of Standardization publishes the list of registered products. Distinctive sign can be placed on the product,

on its package, in leaflets, etc. Other certifying organizations functioning on the basis of MSZ EN 45011 are entitled to carry out tests and grant distinctive signs (differing from those granted by MSZT).

9. Product Warranty and Protection of Customer

The European Community prescribed for member countries in 1985 to include a Product Warranty Law into their juridical system and to enact it according to their respective constitutional means.

In Hungary, there is a Law on Product Warranty [10] in harmony with EU Directives issued on this matter.

A product shall be considered as defective, if it does not provide the expected security, with particular regard to its function, to its reasonably expected use, the information about product properties, the sales date of product, the actual state of science and technology. Product manufacturer shall be liable for the damage caused by the fault of product. In the case of an imported product, the stipulations of Law concerning the manufacturer shall apply to product importer accordingly, who can enforce its claims against the manufacturer.

According to the Law on Customer Protection Main Inspectorates [11], main inspectorate and inspectorates shall inspect, among others, the prescriptions of laws or of authorities concerning the commercialization of products respected, and the prescriptions concerning quality and quality certification of goods and consumer services respected, how complaints for poor quality are settled.

References

[1] Directive No. 89/392 (EEC) of the Council of European Communities from June 14th, 1989 on harmonization of legal rules of member states in relation to machines.
[2] Directive No. 73/23 (EEC) of the Council of European Communities from February 19th, 1973 on harmonization of legal rules in relation to machines involving risks of electric origin.
[3] Hungarian Law XCIII of 1993 on Labor Safety.
[4] Decree No. 5/1993 (Dec. 26th) MüM of the Minister of Labor enacting some stipulations of Law XCIII of 1993 on Labor Safety.
[5] Decision No. 90/683 (EEC) of the Council of European Communities from December 30th, 1990 on different evaluation methods of adequacy defined in technical harmonization directives.
[6] Decision No. 93/465 (EEC) of the Council of European Communities from July 22nd, 1993 on modules applying to adequacy test procedure phases and on CE sign of adequacy (Hungarian Standards Institute, 1995.)
[7] Interpretation of ISO 9001 in the Practice (Hungarian Standards Institute, 1992.)
[8] Hungarian Law XXVIII of 1995 on National Standardization.
[9] Hungarian Law XXIX of 1995 on Accreditation of Laboratories, Certification and Supervisory Organizations.
[10] Hungarian Law X of 1993 on Product Warranty
[11] Hungarian Governmental Decree No. 95/1991 (July 23rd) Korm. on Consumer Protection Main Inspectorate

[12] Hungarian Governmental Decree No. 113/1998 (June 10th) Korm. on Establishing, Operating and Inspecting Lifts
[13] Decree No. 21/1998. (April 17th) IKIM of the Minister of Industry, Tourism and Trade on Safety Requirements and Certification of Adequacy of Machines.

PERIODICA POLYTECHNICA SER. TRANSP. ENG. VOL. 28, NO. 1–2, PP. 65–75 (2000)

TOWARDS SUSTAINABLE MOBILITY – CONFLICTS, IMPACTS, PRINCIPLES, CONDITIONS, ASPECTS OF CEECs

Péter MÉSZÁROS

Department of Transport Operation
Budapest University of Technology and Economics
H–1521 Budapest, Hungary
Fax: 36-1-4633269
Phone: 36-1-4631978
e-mail: meszaros@kku.bme.hu

Received: October 31, 2000

Abstract

There is a growing realisation that the present economic and technological tendencies are unsustainable, in this respect the transport sector came to crossroads, due to its increasing conflicts with the living and built environment, the degradation, loading, pollution impacts.

This study surveys the impact and conflict elements between transport and environment, referring to the main three geographical levels – global, regional and local – including air pollution, land use and intrusion, noise, congestion, accidents and related social impacts. Regarding the basic environmental principles, the study submits proposals for their adaptation in the transport issues, first of all: sustainable development, the precautionary approach, the prevention, the responsibility, the polluter pays principle and other elements, like the integration approach, the environmental limits, the demand management, the environmental efficiency or the equity, as principles.

Concerning the environmental conditions of sustainable mobility, the study surveys the measurable transport impacts including the environmental objectives and selects criteria for sustainability. These criteria are: nitrogen oxides, volatile organic compounds, ozone, particulate matters, carbon dioxide, land use and noise. Regarding air quality, the study proposes vehicle emission and ambient air pollution levels, referring to WHO-EURO levels and OECD proposals and this is the case for land use and noise impacts as well.

Finally the study selects tasks and preferences for the main fields of transport, including international and domestic freight transport, domestic and urban passenger transport and soft modes like cycling and pedestrian mobility.

Keywords: sustainable mobility, environmental impacts of transport.

1. Introduction, Background

Transportation and different forms of mobility are considered in today's societies as essential services, including freight transport with access of goods from production to the markets and passenger transport on the professional mobility and in other social functions. Transport is a significant driving force of the economy, from manufacturing to services, including effects on regional development, employment and competitiveness.

At the same time, there is a world-wide growing perception, with different sectoral emphasis and on global level, that the present tendencies, approaches, technologies are *not sustainable* on a long term run, they will have strong environmental, public health, finally economic consequences.

In this respect transportation, the transport sector has been on cross-roads these days, first of all due to its *increasing conflicts with ecosystems* and built environment, and degradation, polluting, contaminating impacts. The so called transport problem has increasing importance compared to other sectors and their environmental impacts. While in the traditional sectors strong improvements of efficiency, changes of structures have been taking place during the last two decades, in the transport sector in spite of the technical development there is a characteristic expansion, quantitative growth with increasing environmental impacts.

During the last two decades, there has been a 30% increase of transport related CO_2 emission, compared to a 2% decrease of other sectors. In our region a specific problem is the growth of local and regional impact factors, pollutants, so the transport related NO_x and VOCs emissions have grown over 50% and this is over 60% at CO.

Besides air pollution, further impact factors have a growing role in the industrialised countries and in our region, these are: congestion, accidents, noise and vibration effects and the increasing land use, urban sprawl, with loss of biodiversity.

These problems, impact factors have different weight in our regions. The Central-Eastern European countries (CEECs), including those concerned in the EU integration process have a comparatively underdeveloped infrastructure and transport background. In this respect it should be emphasised that the *transport system development* of these countries should be completed on a sustainable approach, avoiding the exhaustion of the natural resources and the repetition of those mistakes committed by the Western-European region's countries during the last decades.

2. Impact Factors, Conflicts between Transport and Environment

Today's transport systems require strong natural resource input, first of all, non-renewable fossils. Beyond the direct adverse effects, the background industries, such as vehicle manufacturing, fuel production, construction of transport infrastructure etc, have additional impacts. The transport emissions appear on three geographical levels, on *global, regional and local scales*. Main global impact contributors of transport are the gases with direct and indirect greenhouse effects, including methane (CH_4), nitrous oxides (N_2O), carbon dioxide (CO_2) and nitrogen oxides (NO_x), volatile organic compounds (VOCs) and carbon monoxide (CO), through the contribution to tropospheric ozone (O_3), as greenhouse gas. NO_x, N_2O and CFCs (chlorofluorocarbons) have stratospheric ozone depletion impacts. CFCs which are still present in the transport sector have even strong greenhouse effects as well.

On regional level the main factors are the acidification pollutants as NO_x ,, SO_2,

and O_3, being strong risk factors in our climate conditions, through the reactions of NO_x and VOCs with public health and ecosystem impacts. Further photochemical oxidant impacts have CO, NO_x, VOCs and SPMs (suspended particulate matters).

Among local effects we should emphasise the toxic impacts of CO and the SPMs through respiratory diseases. VOCs and SPMs especially in associated forms are strongly toxic, even more carcinogen. These elements are characteristic in urban, suburban regions, with cumulation of different pollution factors. Many investigations have proved that in today's urban regions the main air polluting risk factors are the emissions of combustion engines.

As a further effect we should emphasise the *transport related noise*, which is the main contributor to noise impacts. In the industrialised countries 20% of population is exposed to high noise levels, in our case the trunk roads' regions are exposed to 70 dB day-time and 65 dB night impacts, over the WHO limits.

The *land use* of transport is a significant impact factor, in our case in the CEECs being a strong challenge and danger. 5% of the Western part of Germany is covered by transport routes, plus the additional service, infrastructure and background functions. This could be in urban regions more than 25–35%. Additional elements of these tendencies are the sub-urbanisation, the urban sprawl, with use and distruction of active green, agricultural fields endangering the ecosystems and biodiversity. These structures have increasing transport energy use, growing mileage and individual car use.

Congestion belongs to the most adverse effects of transport, reducing the effectiveness of the original functions of vehicles and networks, with time delay, growing fuel consumption and pollution. A further consequence of congestion is the extension of capacities by infrastructure building, itself as an impact factor, inducing additional mobility, attracting traffic flows, reproducing the congestion, inefficiency and environmental impacts.

Accidents are not always considered as environmental impacts, but have a strong contribution to the unsustainable tendencies. Although specific data show certain improvement based on technical and safety measures, but as an indirect effect, there are segregation tendencies discouraging people from walking, cycling. A further element is a social impact in connection with the car dependency marginalizing some social groups, like children, elderly people, disabled and poor population, having limited access to mobility.

3. Environmental Principles, Transport Concerns

Main strategic elements of sustainable development, with reference to transport issues:

- Economic development, structural change *within natural and environmental limits* – in these frames the development programs, plans should be submitted to environmental impact assessment processes.

- During structural change we should prefer *land-saving solutions*, avoiding green field developments. This refers in the field of transport to the neglected, undeveloped existing networks, promoting their rehabilitation, upgrading, revitalisation.
- We should save our *advantages* in the field of environmental conditions, natural resources, even if we have opposite experience from the developed countries. Here we should refer to the safeguarding of our *traditional positions* in the field of public, communal forms of transport, the modal share of rail, the comparatively low level of motorization and motorway network.
- The *environmentally oriented influence* of products within market frames, preferences for environmentally sound products correcting the market value judgement. This refers to the support of vehicles, fuels, technologies with good environmental patterns, by tax, charge preferences promoting public transport, combined freight transport, demand side management solutions.

Basic *environmental principles* to be integrated, to get accepted in the field of transport, including the development process:

- *Precautionary approach*, the minimisation of environmental risks. This is important in those fields where the consequences, impacts are growing with the technical development. This is the case with the transport of dangerous goods but we can refer even to long term effects, like impacts resulting climate change.
- *The prevention* should be a basic economic requirement due to the fact that prevention is much less expensive than subsequent measures. In this respect those products, technologies should be supported, promoted, which have less environmental impact, resource need. We refer here to methods of transport telematics, application of mobility management methods or even development of bicycle route networks.
- In the frame of *partnership* , the local actors, municipalities, citizens' groups, local business community representatives should have an increasing role in the realization of environmental objectives. This could appear in local decision making on local developments, even businesses. In the field of transport this could cover several initiatives, from traffic damping to logistic center projects.
- The *host-approach* refers to that fact that environmental tasks, issues are not purely economic or market oriented activities, but they could be 'only' safeguarding or creating values, certain exemplary and responsible behaviour. This is necessary at local municipalities, authorities, having positive impacts on local population, entrepreneurs, contributing to the social participation. This could be manifested on local level in initiatives to bicycle routes, local public transport, or other positive developments.
- The *responsibility* or the polluter pays principle is an objective fundamentally debated in documents all over the world. Nevertheless the payment for the use of resources should be a particular objective. In the field of transport, all

internal and external costs of the use of environment should be paid by the user, including the damages caused in natural resources, human health, built environment. Enforcement of these costs is not a pure financial, economic issue, but it should be accompanied by additional push and pull measures, restrictions, regulations.

- The *planning principle* requires long term strategies in the field of environment. The two sides of this issue could be the National Environmental Programs with direct, indirect transport relations, outlooks, continuous regional and local, municipal environmental programs. The other side should be the Transport Policy, or Transport Master Plan, with regional, local levels, connections with regional development plans, integrating the environmental approaches.
- The *integrated pollution control* is a certain extension, continuation of the above issues, requiring the integration of environmental approach into other social, economic processes, so in the field of transport and mobility. On the other hand it requires the integrated approach in the field of safeguarding of ecosystem elements, in terms of tools and legal systems. Here we should refer to the institutions of EIA, SEA and Environmental Audit, having strong potentials in the field of sustainable transport development.

Further principles of sustainability, referring to the complexity of this issue, having social, economic, moral, political elements:

- principle of environmental limits,
- demand side management, with mobility management
- environmental effectiveness, with improved resource use
- the welfare efficiency
- the equity

4. Sustainable Mobility – Environmental Conditions

After basic approaches, definitions of sustainable development, including the Brundtland report, the UNCED declaration or Herman Daly's considerations, having common elements on the carrying capacity of the ecosystem and input-output models of energy and resource use, we have witnessed some attempts to define environmentally sustainable transport or mobility during the recent years. Kageson's approach is connected to the basic definitions of sustainability: Sustainable transport's objective is to offer a *basic mobility* to all citizens, without endangering, contaminating the nature and environment. It needs certain clarification of basic mobility needs, expected natural, environmental conditions, current public health risks and pollution limits, with additional timetable, action plans with medium and long term objectives.

The Centre for Sustainable Transport's approach is: Basic access needs of individuals and societies to be met safely and in a manner consistent with human

ecosystem's health and with equity within and between generations, is affordable, operates efficiently, offers choice of transport modes and supports a vibrant economy, limits emissions and waste within the planet's ability to absorb them, minimises consumption of non-renewable resources, reuses and recycles its components and minimises the use of land and production of noise. Beside direct environmental respects sustainability should have *social and economic elements*, reflecting fairness, equity and practical feasibility. It means that sustainable mobility could be combined with qualitative and quantitative criteria, arranging mobility objectives and environmental limits, criteria.

These quantitative factors could express emissions and other impact effects of transport, defining the cumulative character of transport loads. We should refer to the whole verticum of transport, with the background industries, manufacture and maintenance processes, including infrastructure development and waste disposal.

To the survey of impacts and selection of criteria we can use the matrix format arranging the *objectives and domains*, geographical levels.

	Types of objective		
Domains	**Public Health**	**Ecosystem Quality**	**Resource Management**
Global	Stratospheric O_3 Persistent compounds	Greenhouse gases Statospheric O_3 Biodiversity	Energy use Material use Recyclability
Regional	Troposphere O_3 Persistent compounds	Troposphere O_3 Acidification (NO_x, SO_2) Persistent compounds Nitrogen deposition	Land use Energy use Waste
Local	O_3, VOCs, PM Carcinogens Individual health and quality of life Noise, Safety	Impacts on the urban ecosystem, incl. landscape and separation of functions by transport and its infrastructure	Damage to the buildings Energy use Land use Waste

Beside these elements we can raise the level of mobility, but in this stage the main issue is the consideration of environmental conditions. In analysis of impact factors, we would give priority to the *public health* in strong correlation to the *protection of natural resources*. Based on these considerations, the following *criteria* could be proposed for construction of environmental conditions of transport:

- Nitrogen oxides (NO_x)
- Volatile organic compounds (VOCs)
- Suspended particulate matters (SPM)
- Fossil Carbon dioxide (CO_2)
- Land use in urban, sub-urban regions
- Noise impacts

These criteria mostly cover the collected and surveyed impact and risk factors of the table. Nitrogen oxides are good marking elements of combustion engine processes, having at the same time various impacts on all levels. VOC emissions are characterising the fuel verticum, while SPMs are the strongest risk factors of urban air pollution in connection, first of all, with diesel engines. The fossil CO_2 is another basic indicator of energy use and greenhouse effects. The transport related land use is an important factor expressing numerous impacts, like the paving of lands, urban sprawl, suburbanisation, adverse effects on biodiversity. Noise is in direct connection with motorised transport and its role, its contribution is over the air pollution in Europe.

Beside those elements we could mention the accidents and waste disposal as transport externalities, negative impacts, emphasising the lifecycle approach, but the above listed criteria have a good coverage of transport related impacts.

5. Analysis and Quantitative Proposals for the Criteria

- *NO_x*. Transport's share in NO_x emission is 60% in Europe, over 50% in Hungary. The ambient air pollution in several regions of Europe is over the critical limit, contaminating the ecosystem by acidification and eutrofisation. In spite of the increasing role of catalytic converters there is an expected growth of emission in non OECD countries by 40% and an estimated fall in OECD regions by 20% until 2030. Many experts claim a 90% decrease of emission in correlation with a 30 $\mu g/m^3$ ambient level (annual mean) as a sustainable impact in urban and suburban regions. The 90% fall of emissions could be realistic even within today's technologies, with additional improving fuel efficiency. Regarding emissions, current data of Otto engines of 0.6–0.7 g/km could be reduced to *0.12 g/km* which is technically feasible, being as sustainable criterion in this respect. For heavy, diesel engines the current EU level of 0.7–0.9 g/t.km could be reduced by 70% to *0.2–0.25 g/t.km*, which is a feasible level for sustainability. Due to the strong potential on fuel efficiency, the general 90% emission reduction should be realistic, targeting a *40 $\mu g/m^3$ limit* for urban conditions.
- *VOCs*. Many of them have direct toxic effect on humans and ecosystems, first of all, benzene and butadiene have carcinogen effects, while most of them have greenhouse and photochemical impacts. Main sources of emission are gasoline engines, through evaporation and unburned fuels. These elements raise technical tasks, improvements on engine design, fuel and exhaust systems. From technical point of view both for engines and fuel systems there are strong potentials, which could lead to a complete elimination of emissions. Levels of sustainability could be proposed as WHO/EURO impact levels: *benzene: 2.0 $\mu g/m^3$, butadiene 0.1 $\mu g/m^3$, PAH 0.5 $\mu g/m^3$*.
- *Ozone*. Needs special attention, being a direct public health factor. Based on international figures, critical concentration is above four times the harmful level. There is a certain relation between ozone formation and NO_x and VOC

level, but at catalytic converters, removing NO_x, there could be increasing ozone levels. On a long term level, the reduction of ozone concentration is the main public health and ecosystem challenge in strong correlation with NO_x and VOC control. Level of sustainability could be on residential areas *120 $\mu g/m^3$* (8 hours ambient level).

- *Particulate matter.* In certain opinions, SPMs represent the most serious urban risk factor these days, although in concrete atmosphere only a certain amount could be attributed to motor vehicles, first of all, to diesel engines. From this category the diameters of less than *2.5 μm* have a paricular risk to respiratory systems, associated with hydrocarbons or benzapyren as strong carcinogen substances. Reduction of this impact could be achieved by low sulphur level fuels and better combustion processes. The related sustainable ambient impact level, based on WHO/EURO orientations could be *15–20 $\mu g/m^3$* (annual mean) referring back to the high risk of particles less than 2.5 μm.
- *Fossil CO_2.* One basic criterion of environmentally sustainable transport should be the control of CO_2 emissions. The IPCC panel proposed a general reduction of 50–70% on a short term basis, contributing to the stabilisation of atmospheric concentration. Certain experts argued for a 80% reduction of emission in the industrialised countries, while we can witness a permanent growth of transport related CO_2 emission all over the world. We should note that there are crucial differences between specific emissions of different regions and countries: USA 5.8, Germany 2.3, Hungary 0.7 tonnes CO_2/capita/year and between modes of transport. Regarding reduction opportunities, there are short term potentials of 60–80% in energy intensity of vehicles, targeting vehicle construction and performance, which could be extended by the introduction of renewable fuels, positive corrections of modal split. In this respect there are certain regulations in the USA for ZEV (zero emission vehicles), and introduction of non fossil vehicles would be necessary. Criterion for sustainability in this issue should be *70–80% reduction* of transport related CO_2, first of all, in the OECD countries. In spite of our region's relatively favourable position we should make steps in this respect, in connection with the economic growth and its consequences of increasing CO_2 emissions.
- *Land use.* Transport related land use is a particular adverse effect of urban, suburban regions, degrading ecosystems. Here we should refer to urban sprawl, suburbanisation processes, having increasing mobility, first of all, car use demands beside loss of active green fields. Transport related land use in developed regions is over 20–25%. As quantitative criterion for transport purposes, including maintenance and other background infrastructure a *10% limit* could be determined, accompanied by further measures, for safeguarding, rehabilitation and limited disturbance of living ecosystems.
- *Noise.* Transport related noise became a significant public health risk, with different direct and indirect adverse impacts – stress related psychiatric illnesses, reduced immunity, blood pressure. In our region there is a charac-

teristic noise impact of 70–75 dB by day and 60–70 dB at night, along main roads. WHO proposals refer to *55 dB day-time and 45 dB night* levels in residential and 50 dB/40 dB in recreation areas, being acceptable from the point of view of sustainability.

6. Related Tasks, Preferences

Besides technical measures regarding emission control of gases mentioned among the criteria and other emissions, there is a need for a strong *policy change* both in the field of transport development, management of transport systems and additional regulative policies. Main elements of these measures in some transport sub-sectors:

- *International/transit freight transport.* A structural shift of freight flows towards combined transport. Additional development of transit charges and regulative systems, in connection with vehicles' environmental performances. Promotion of development of vehicle pools of combined transport targeting rail and waterway technologies, including the necessary logistics chains. Formulation of transit corridors, avoiding sensitive areas.
- *Domestic freight transport.* Structural development towards logistics principles, considered development of logistic centers. Application of differentiated vehicle charges, relating to environmental performances, with green card system for lorries. Extension of time, regional and weight limitations.
- *Domestic interurban passenger transport.* Improvement of rail transport's conformity, access time – competitiveness, increase of speed with upgrading of rail network, special vehicle pool for small rail passenger flows. Integration of transport modes.
- *Individual car use.* Car users should be charged by environmental impact related, proportional fees, costs. A network of development, categories of environmental impact assessment should be introduced, including service and management facilities. Improved control system for vehicles' technical, environmental registration and control.
- *Local passenger transport.* In urban road networks, the tolerable load, environmental impacts should be the determining factors instead of mobility demands. Preservation and improvement of relative price positions of public transport. Introduction of road price, cordon systems in urban areas, traffic priorities for public transport vehicles. Integration of rail networks in local and suburban public transport. Further mobility demand management in sensitive urban regions with growing car free areas.
- *Bicycle mobility.* Extension of networks, with additional serving infrastructure – parking, lending, repair etc. Integration with public transport, regulation of safety measures and related traffic.
- *Pedestrian mobility.* Strong potential in urban regions by car traffic control, improvement of ratio of pedestrian areas. Measures, programs towards access, safeguarding of pathways from other functions.

7. Central-Eastern European Aspects, Priorities

During the recent year, a multi-country Phare Project has been launched in the field of transport and environment, covering the 13 Phare countries of the Central-Eastern European region. Main objectives were, targeting a multi-country approach, the reduction of environmental impacts from the expected growth of transport, and integration of environmental issues into transport policies. After the review of current situation of the field of transport and environment in the concerned countries, there has been a strategy formulation and in this period an action planning has been taking place. Main elements, experience of the *review* raising the most crucial problems of the region include:

- Outdated and old vehicle fleet both in road and rail transport, with high emissions, low fuel efficiency and noise impact.
- Increasing motorization with insufficient infrastructure, particular in urban, suburban areas.
- Low fuel quality with environmental consequences.
- Expected increase, pressure of road infrastructure development.
- Loads, contaminations of recent and former activities, including waste disposal, fuel storage and vehicle management.

The formulation of *key objectives* was based on the current problems, existing and planned policies, regulations of the concerned countries and international regulations, policies relating to transport and environment. The objectives can be classified in three main categories:

- reduction of emissions
- transport demand management, modal split influence
- improvement of environmental performance of transport infrastructure

Key objectives:

7. To enhance the use of cleaner, more economical and quiet vehicles.
8. To enhance the use of cleaner fuels.
9. To reduce the car mobility and to encourage road transport efficiency.
10. To reduce the noise nuisance along main roads.
11. To promote transport by rail.
12. To enhance the use of maritime and inland waterway transport.
13. To strive for sustainable development of air transport.
14. To improve urban traffic management.
15. To improve the environmental performance of existing and new transport infrastructure.
16. To optimise combined transport.
17. To enhance the safety of transport of dangerous goods.

Some of the *proposed measures and instruments* contributing to the above key objectives:

1. Financial and taxation instruments.
2. Promotion of the use of cleaner and alternative fuels.
3. Variabilisation of external costs.
4. Setting noise emission standards and implementation of economic instruments.
5. Improvement of the railway infrastructure management.
6. Improving adaptability to intermodal transport systems.
7. Implementation of fair and transparent pricing.
8. Discouraging and control of motor vehicle traffic in densely populated areas.
9. Strengthening transport-environmental policies.
10. National plans and programs for implementation of combined transport.
11. Implementation of national control and reporting systems.

In the final and current stage, action planning is taking place, targeting the implementation of the key objectives on multi-country and national level.

References

[1] Commission of the European Communities: Progress report from the Commission, on the implementation of the European Community Programme of Policy and Action in relation to the Environment and Sustainable Development. 'Towards Sustainability' Brussels 10.01.1996.
[2] Pollution Prevention Control. Environmental Criteria for Sustainable Transport. Report on Phase 1 of the Project on Environmentally Sustainable Transport. OECD. Paris 1996.
[3] HOLMAN, C.: The Greening of Freight Transport in Europe. European Federation for Transport and Environment. Brussels. 1996.
[4] European Sustainable Cities. Report of the Experts Group on the Urban Environment. European Commission. Directorate General XI. Brussels 1996.
[5] The Citizens' Network. Fulfilling the Potential of Public Passenger Transport in Europe. European Commission Green Paper. Brussels Luxembourg 1996.
[6] Sustainable Transportation: Survival in the 21st Century. Rod Diridon, Norman Y.Mineta, International Institute for Surface Transportation Policy Studies. San Jose State University. 1998.
[7] KAGESON, P.: The Concept of Sustainable Transport. T&E Brussels, 1994.
[8] Transport and the Environment, A Multi-Country Approach. Second Interim Report, Strategy Formulation. 1999. European Commission, DHV, LT Consultants.
[9] National Environmental Program. Ministry of Environment. Budapest. 1996.
[10] Transport Policy of the Government of Hungary. Budapest 1995.
[11] Environmentally Oriented Transport Policy Concept. Hungarian Traffic Club. 1994.
[12] Transport and the Environment: A Multi-Country Approach. Strategy Formulation. Phare European Commission. DHV, LT Consultants. May 1999.

PERIODICA POLYTECHNICA SER. TRANSP. ENG. VOL. 28, NO. 1–2, PP. 77–90 (2000)

NUMERICAL MODELLING OF THE MECHANICAL PROCESS OF PARTICLE DETACHMENT BY FINITE ELEMENT METHOD

András ELEŐD*, János DEVECZ* and Tibor BALOGH**

*Department of Vehicle Parts and Drives
Budapest University of Technology and Economics
H–1521 Budapest, Hungary
**Széchenyi István College
H–9026 Győr, Hungary

Received: Oct.10, 2000

Abstract

This paper presents a numerical simulation technique for calculating the strain and the stress state at every point of the surfaces, containing real surface topology and layered structure. Depending on the nature of deformation, we can evaluate the actual stress and strain state either from the point of view of the fatigue failure or from the point of view of the exhaustion of deformation capacity of the surfaces. Using the stress and strain components calculated by the Finite Element Method at every nodal point of the surfaces, it is possible to predict the probable surfaces degradation mechanism, i.e. to predict the probable mechanism of the particle detachment under dry sliding conditions, under consideration of the nominal load and to the required lifetime of the surface.

Keywords: tribology, numerical simulation, finite element method.

1. Introduction

Several researchers have been dealing with the possible forms, mechanisms and quantitative determination of wear developing from friction since the beginning of the century. These theories have a common feature, they do not differentiate between the mechanisms of initial wear and permanent wear and neither of them have satisfactory explanation for friction without wear.

In the recent decade, the 'third body' theory, the investigation and explanation of friction from a new viewpoint has come forward. Based on the results of the tests that are given by modern test machines, the 'third body' theory queries the usability and validity of the classical wear theories or wear mechanisms in some aspects. The 'third body' theory says that the friction mechanisms known so far explain only initial wear and not wear in the permanent stage of friction [1]. In the permanent stage of friction the particle detachment means the so-called inside source of the development of the 'third body' and only particles leaving the tribological system permanently are regarded as wear [2]. The wear is a result of a complex process, but the stipulation of the wear beginning is, in any case, the beginning of the particle detachment.

2. Possible Forms and Conditions of Particle Detachment of Metals

Particle detachment is the result of fracture occurring when the deformation capacity of the material is exhausted. Except for the extreme case of micro-cutting, the particle detachment can happen practically in three different ways:

- *in form of brittle fracture*: as a result of elastic stress state, without starting the plastic deformation, in consequence of a specific fracture work that is reduced to zero under influences of state-factors. In the case of centred metals of cubic lattice with the most sliding planes, fracture without plastic deformation is not expected to start. In the case of space centred metals of cubic lattice and even more in the case of metals of hexagonal lattice having one sliding plane, brittle fracture may occur following the joint effect of state factors. The particles approaching the brittle state at the moment friction begins and with their specific fracture work becoming zero, will fracture without plastic deformation. The detachment of these particles means the so-called initial wear experienced at the beginning of friction.
- *in form of fatigue fracture*: as a consequence of the cumulative damage, under elastic contact conditions. In the case of stress smaller than the yield point, taking the accumulation of damages into account, fracture occurs when the aggregate of the work of the elementary plastic deformations reaches all of the work necessary for fracture.
- *in form of ductile fracture*: when the strain of the given particle during deformation exceeds the critical limit values, i.e. the deformation capacity characteristic of the material of given state.

From the possible procedures that are traceable to mechanical causes of the particle detachment, it is possible to indicate in advance the development of the fatigue fracture and the ductile fracture. By the help of the Finite Element Method, we can determine at every nodal point of the model the components of the stress and strain states, in function of the external load and of the geometry of sliding surfaces.

In the following sections, we would like to show, with the help of FEM-simulations of dry friction, how we could predict the expected form of the surface degradation for a given surface topology and material properties.

3. Real Surface Contact Model

The numerical modelling of contact problems – due to the development of the computing technology – suffered a significant change during the last years. From the linear-elastic 2D solution to the non-linear 3D elastic-plastic solution every step of the development can be followed in the literature. However, the latest studies in this field are restricted either to the two-dimensional analytical solution of the elastic-plastic problem [3], [4], [5] or to the approximate solution of the 3D

contact [6], [7], [8], [9]. Authors using the approximate three- dimensional solutions assume frequently elastic perfectly-plastic material behaviour and/or axisymmetric coordinate system.

Our numerical simulation is based on the solutions by Finite Element Method of the three-dimensional contact of surfaces which is characterised by non-linear hardening law and by real topography. FEM is preferred for the contact simulation, because it can place in our disposal all components of the strain and stress states at every moment of the movement and at every grid point on the surfaces.

To simulate 3D contact, first it is necessary to develop the geometrical model of the surfaces. For developing rough surfaces, there are basically two options:

- to generate a surface roughness with the help of individual roughness contours [10], [11] [6], [12] or with specified auto-correlation function and distribution function of roughness peaks [13], [14], [15],
- to use the measured surface topography in discretized form [7], [9], or in digitalized form.

In the second case, the discretization signifies the connecting of the measured roughness values by line sections in both directions, so the surface will be covered by elementary squares. The digitalization of the measured surface means, that the measured roughness values are introduced directly into a CAD-software where a so-called free-form surface can be generated.

Our geometrical model consists of two solid bodies with real surface topography and with layer structure, putting one on top of the other. The surfaces we have developed and used for our simulations are free-form surfaces, based on measured roughness values of a polished surface area of 1 mm^2. The density of registered points was 1 μm in both directions. The generated mathematical surface, containing all measured roughness value, is presented in *Fig. 1*/a.

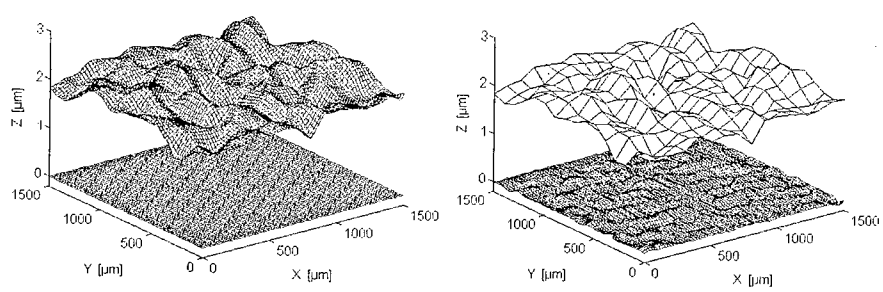


Fig. 1. The measured (1/a.) and the approximate (1/b.) surface topology

To determine the minimum needed number of grid points of the meshed surface, a comparative calculation was made between the mathematical (free-form) surface containing all measured points and another one containing a reduced number of points. This second surface was subtracted from the original one and as long as the difference-surface was near to plane, we considered that the surface which was

generated by a reduced number of points characterises satisfactorily the original one (*Fig. 1*). In this way, using the mesh resolution represented in *Fig. 2*, we could take into consideration approximately 20 roughness peaks on a surface area of 160×300 μm. The question is whether the results obtained by this surface area are generally valid for isotropic surfaces.

For the pre- and post-processing of the numerical contact model (*Fig. 2*) and to solve the numerical problem by the Finite Element Method, we used the FEA Lusas 12 software with Mystro pre- and post-processor. The mesh was generated using HX8M elements. To define the contact conditions between two solids, we used slide-surface pre-contact facility without friction.

The material of the lower solid body is C55 steel with elastic-plastic material properties, the upper one was considered as an ideal rigid body. Elastic properties are determined by Young's Modulus ($E = 2.1 \cdot 10^5$ MPa) and by Poisson's Ratio ($\nu = 0.3$). Plasticity is characterised by the flow curve of the material. Former results of FEM calculations show that the sign of the stress components changed before and after the contact of the asperities, i.e. the asperities have an alternate compression and tension stress state during the friction.

However, flow curves are generally determined under simple stress conditions, i.e. uniaxial tension or compression conditions.

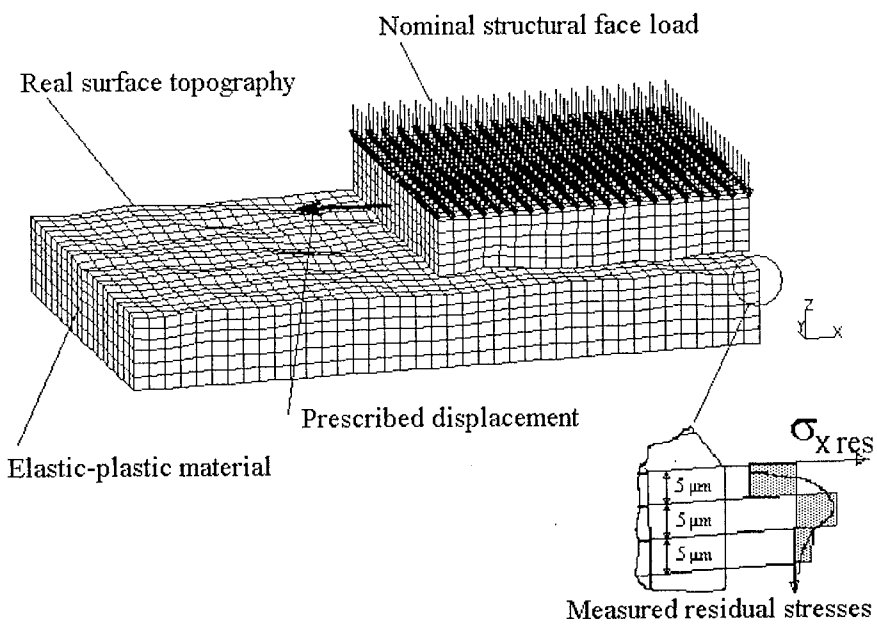


Fig. 2. The numerical contact model

To better characterise the plastic behaviour of the surface, we used the so-called cyclic flow curve (*Fig. 3*), which was put under alternate compression and

tension stress conditions. The approximate power strain hardening law of the cyclic flow curve is given by $\overline{\sigma} = 812 \cdot (\overline{\varepsilon}^{pl})^{0.15}$[MPa].

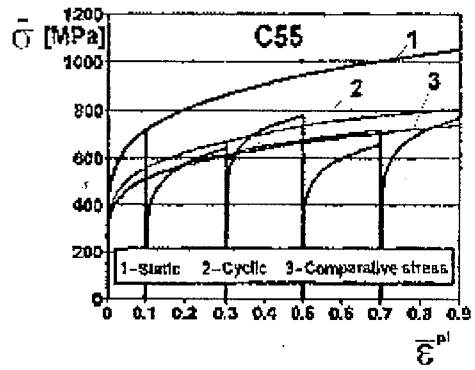


Fig. 3. The cyclic flow curve of the C55 steel [16]

The loading of the surfaces consists, on the one side, of the structural face load (50 MPa), on the other side of the rigid body-like displacement (totally 150 μm and step by step 1 μm) of the upper solid to the direction of global x axis. The elastic half-space of the lower solid was taken into consideration by elastic supports. The degree of freedom of these supports was subjected to specified spring stiffness equal to the Young's Modulus of the steels.

At the tribological processes, the existence of residual stresses developed during surface machining in the near-surface layers has a significant influence on the development of actual stress state. The residual stresses regarding their feature may be regarded as σ_x or σ_y stress component developing on the plane of the surface. If the stress state, when developing, meets the yield condition, local plastic deformation is carried out and in this way the surface layer can be locally hardened. The important influence of the steady state residual stresses on the fatigue crack initiation is explained by [17]. Following probable local plastic deformation their fraction not exceeding yield condition remains in the surface layer in measurable form and is added to the stresses developing during friction with sign. In this way, the residual compressive stresses decrease equivalent stress, in the favourable case, may only create hydrostatic compressive stress resulting in elastic deformation while the residual tensile stresses increase the brittle tendency.

In our contact model (see *Fig. 2*) the residual stress state of the polished surfaces, measured by an X-ray diffractometer, was also considered:

Surface/Depth	0...5 μm	5...10 μm	10...15 μm
polished	$\sigma_{x\ \mathrm{res}} = -119$ MPa	$\sigma_{x\ \mathrm{res}} = -103$ MPa	$\sigma_{x\ \mathrm{res}} = 76$ MPa

To evaluate the results of the numerical analysis, we calculated the average principal strain and stress components at every nodal point, as well as the equivalent strains (in the figures: EE = Epsilon Equivalent and EPE = Epsilon Plastic Equivalent) and the equivalent stresses (SE = Sigma Equivalent). Directions of the principal strains are consistent as usual with the directions of the principal stresses and are based on the maximum value.

During the simulation, first one slide-cycle will be examined and the character of the deformation (elastic or plastic) will be observed by the evaluation of the calculated results of the numerical analysis. In the case of the elastic deformation, the nodal results will be evaluated from the point of view of fatigue of the surface (see Section 4). By the appearance of a localised plastic flow (deformation plastic), the actual deformation capacity of the given nodal points of the surface has to be considered (see Section 5) and the cycle of movement has to be repeated, on one hand, until the steady state of elastic shakedown limit is reached, on the other hand, until the exhaustion of the deformation capacity, i.e. until ductile fracture.

4. Numerical Prediction of Particle Detachment for the Case of Elastic Deformation of the Surface

If the stresses, which developed in the surface layers during one cycle of sliding, remain only elastic, then the particle detachment may be a consequence of a fatigue fracture.

Former studies traced the fatigue failure of the surface back to the fatigue of the individual roughness peaks. That was the mistake of KRAGELSKI's model [18] too. ZHARIN and co-authors [19], based on their observations through a microscope, explained the fatigue failure of the surfaces with fatigue fractures in the subsurface layers. Recent studies try to find new criteria of rupture to explain the failure under repeated cyclic stress in the case of elastic-plastic surfaces. French authors prefer to use the Dang Van criterion to predict the failure of the surface. Dang Van criterion takes into consideration separately the influence of normal and shear stress components, stressing hereby the role of the accumulation of the plastic strains due to the share stresses [8]. From the point of view of the mechanism of the failure, KAPOOR [20] and JOHNSON [21] separate the strain into two components, to a reversing (fatigue) component and to a unidirectional (ratchetting) component. The hypothesis was then made that two modes of failure can distinguish low cycle fatigue and ratchetting failure.

In the case of our numerical model, the FEM was used to predict the failure under condition of elastic stress state of the surface. The FEM offers the opportunity of giving the actual value of the damage (or the predicted lifetime) automatically in every nodal point as a function of the local equivalent stress, the loading spectrum and the fatigue-curve ($S - N$ curve) of the surface.

The fatigue behaviour of the surface layers of our model was characterised by the $S - N$ curve of the basic material (*Fig. 4/*a). To determine the loading spectrum,

we assumed that:

- during one sliding cycle the number of the theoretically possible contacts is equal to the quotient of sliding path (150 μm) and the double of the mesh parameter (2×7.5 μm),
- during one cycle all roughness peaks are loaded once by 1/3 of the total load (to have a balanced state of solid bodies, 3 contacts are needed), otherwise they are loaded proportionally to the number of the possible contacts,
- under dry friction conditions, only interrupted moving may be possible, so the total friction length is limited by 15 m which yields 10^5 cycles.

The loading spectrum is shown in *Fig. 4*/b.

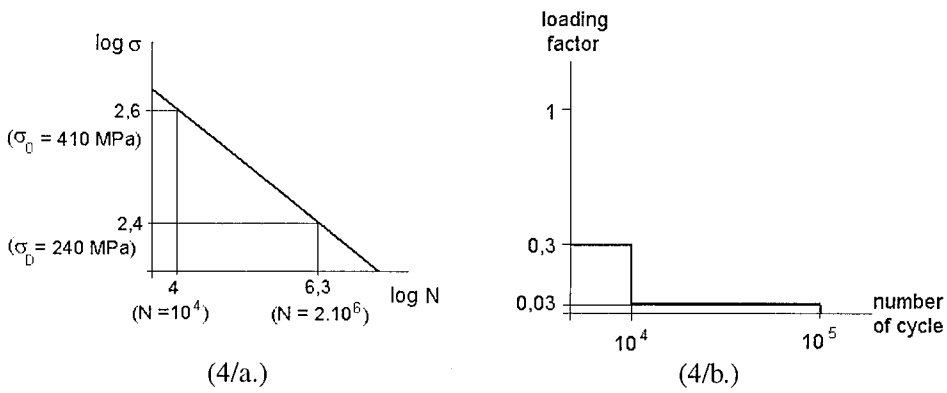


Fig. 4. The fatigue curve of the basic material and the loading spectrum of the surface

Assuming the fatigue behaviour of the surface to be characterised by the $S-N$ curve of the basic material and using the Miner's rule, with the help of the loading spectrum, we can determine at every nodal point of the surface the total of the cumulated damage. If this value reaches one, we can assume that in this region the process of the fatigue failure may start.

In the case of the illustrated model, up to the 59th calculation step we could detect only elastic deformation in the nodal points of the surface, i.e. the equivalent stresses were smaller than the yield stress (*Fig. 5*), but at the 59th calculation step the cumulated damage exceeded the critical value (*Fig. 6*).

If the deformation state of the surface stays on the elastic field during one cycle of the moving, this simplified calculation provides the opportunity to determine the maximum admissible load as a function of the required lifetime of the surface without particle detachment.

In the case of the illustrated model, the first plastic deformation appeared at the 60th calculation step (see *Fig. 10*), so these calculated results had to be analysed in terms of ductile fracture conditions.

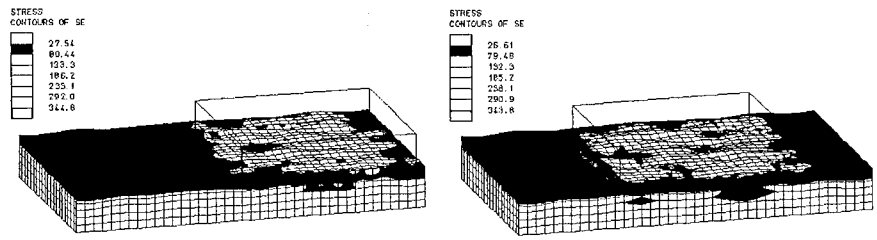


Fig. 5. Equivalent stresses (SE) in the 3rd step after the start (left) and after the beginning of the particle detachment by fatigue fracture (the values of SE are given in MPa)

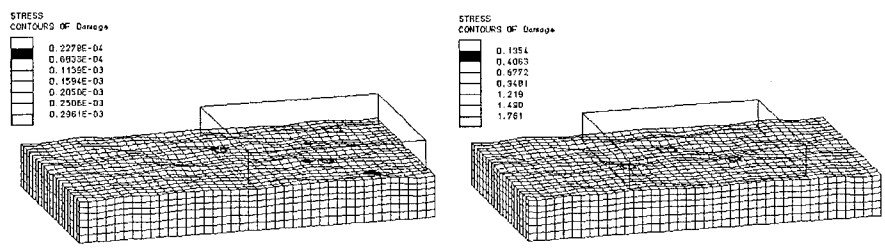


Fig. 6. The cumulated damage at the 3rd step after the start of the calculations (left) and at the beginning of particle detachment (right)

5. Numerical Prediction of Particle Detachment for the Case of Plastic Deformation of the Surface

If the plastic deformation starts at any nodal point of the surface, the mechanical process of particle detachment will be the ductile fracture. The ductile fracture will happen, if the local deformation capacity of the material exhausts for some reason. To define the conditions of particle detachment, first the deformation capacity of the given particle must be known.

Former studies, such as FLEISCHER [22], show the connection of the deformation capacity to the specific fracture work of the material. The specific fracture work (W) is a material testing index, which is equal to the work concerning a volume unit, carried out by the maximal tensile stress (σ^*), until the appearance of the ductile fracture:

$$\int_0^{\bar{\varepsilon}_f} \sigma^* \, \mathrm{d}\varepsilon = W. \tag{1}$$

In *Eq.* (1), $\bar{\varepsilon}_f$ represents the maximal strain without having a ductile fracture, i.e. the deformability or the deformation capacity of the material.

The specific fracture work can be defined by a simple tensile test, as the sum of the work of elastic deformation, the work of uniaxial (linear) plastic deformation

and the work of contraction:

$$W = W_e + W_p + W_c, \tag{2}$$

where the work of elastic deformation is:

$$W_e = \frac{\sigma_0^2}{2E}. \tag{3}$$

In this relation σ_0 is the yield stress, E is the Young's Modulus.

The work of plastic deformation to the limit of uniaxial (linear) strain (ε_m) is:

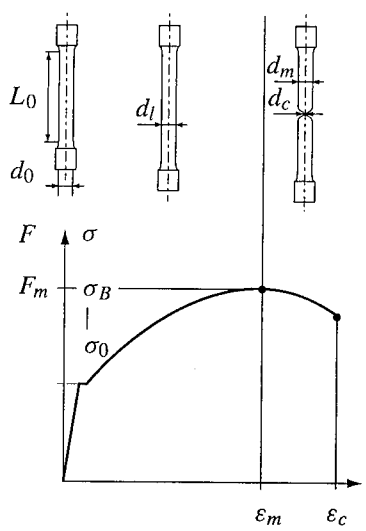
$$W_p = \frac{(\sigma_0 + 2\sigma_B)\varepsilon_m}{3}. \tag{4}$$

In this relation σ_B is the tensile strength.

The work performed during contraction is:

$$W_c = \sigma_m' \frac{\psi_c - \psi_m}{1 - \psi_m}. \tag{5}$$

In the relation σ_m' is the actual stress belonging to the maximum tensile load, Ψ_c is the contraction belonging to fracture, Ψ_m is the contraction at the end of linear strain. With the signs of *Fig. 7*:



$$\sigma_m = \frac{4F_m}{d_1^2\pi}$$

$$\varepsilon_m = \frac{d_0^2 - d_m^2}{d_0^2} \tag{6}$$

$$\varepsilon_c = 1 - \frac{1}{1 + \varepsilon_c}$$

Fig. 7. Determination of the contraction work by tensile test

In our case, the specific fracture work of the examined C55 steel is 443 MJm^{-3}.

The specific fracture work in principle is an index insensitive to the structure, which changes value only if the lattice structure of the metal changes. The stresses used to define the specific deformation work and indexes of strain character, however, are sensitive to structure. They depend on the brittleness or plastic states of the material. Brittleness or plasticity are not features of the material, they only indicate their state. Therefore, brittleness or plastic state of the material depends, in addition to the structure of the material, on the state factors, i.e. on the temperature, on the speed and on the stress state. The effect of the temperature and speed could not be taken into consideration yet for the present work.

However, ZIAJA [23] verified by theoretical and experimental investigations that the specific fracture work is also dependent on the stress state. Assuming a usual plane stress state at the environment of the contact area, the dependency can be described by the following relation:

$$\int_0^{\overline{\varepsilon}_f} \sigma_1 \, \mathrm{d}\overline{\varepsilon} = \int_0^{\overline{\varepsilon}_f} \frac{\overline{\sigma}}{\sqrt{3}} \frac{2+\alpha}{\sqrt{1+\alpha+\alpha^2}} \, \mathrm{d}\overline{\varepsilon} = W. \tag{7}$$

In this relation, $\overline{\sigma}$ is the flow stress as a function of the equivalent plastic strain ($\overline{\varepsilon}^{pl}$) and α is a factor describing the deformation history. The interpretation range of α is in this case:

$$-1/2 < \alpha = \mathrm{d}\varepsilon_2/\mathrm{d}\varepsilon_1 < 1. \tag{8}$$

For the case of the strain hardening power law, the flow stress is defined as $\overline{\sigma} = K \cdot (\overline{\varepsilon}^{pl})^n$, the integration can be carried out in closed form, and the deformation capacity ($\overline{\varepsilon}_f$) as a function of the actual stress state can be defined as:

$$\overline{\varepsilon}_f = \left[\frac{W(1+n)}{K} \frac{\sqrt{3(1+\alpha+\alpha^2)}}{2+\alpha} \right]^{\frac{1}{1+n}}. \tag{9}$$

In the case of our material, the value of the deformation capacity can vary between 0.59 and 0.67 in the given interpretation range.

To characterise the deformation state of the particles of the surface depending on the actual local stress state, a reserve of plasticity (RP) was calculated at every nodal point:

$$RP = 1 - \frac{EPE}{\overline{\varepsilon}_f}. \tag{10}$$

Calculating the value of $\alpha = \mathrm{d}\varepsilon_2/\mathrm{d}\varepsilon_1 \cong \varepsilon_2^{pl}/\varepsilon_1^{pl}$ continuously during the numerical analysis and using *Eq.* (9), it is possible to determine the deformation capacity step by step at every nodal point. If this reserve of plasticity exhausts, i.e. the equivalent plastic strain (EPE) attains the deformation capacity ($\overline{\varepsilon}_f$), the reserve of plasticity will be zero at the given nodal point and the particle detachment, as a result of a ductile fracture, may happen.

The calculated results were drawn in two positions in both models just before the critical deformation state and one step after it:

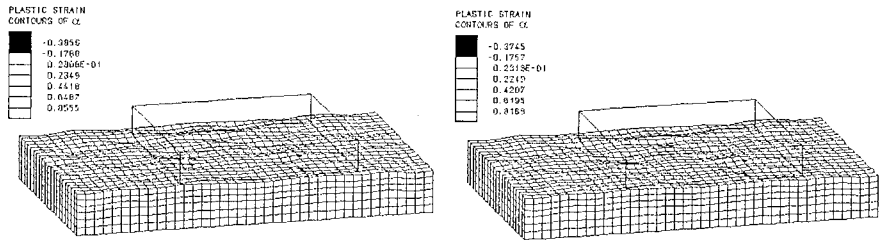


Fig. 8. The variation of α before (left) and after (right) the critical deformation state

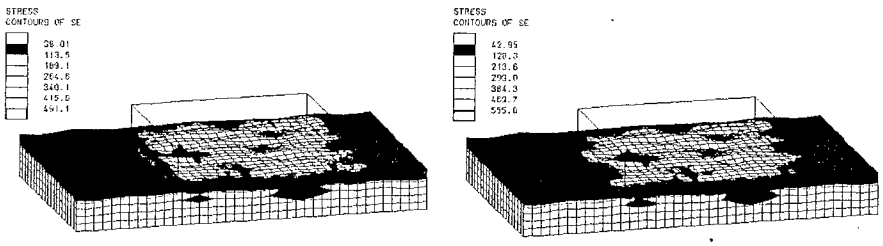


Fig. 9. Equivalent stresses before (left) and after (right) the critical state causing the particle detachment

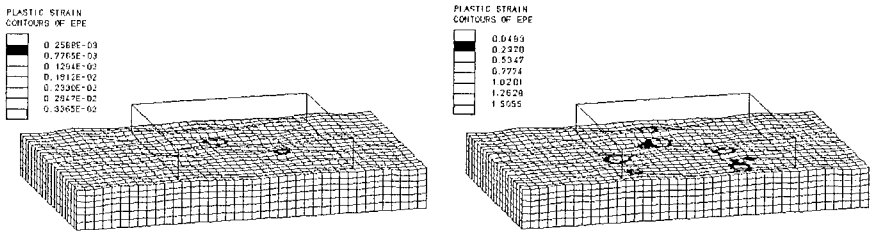


Fig. 10. Equivalent plastic strains (EPE) before (left) and after (right) the critical deformation state

6. Results and Conclusions

Today's computing techniques allow to analyse the three-dimensional contact problems with the help of the Finite Element Method. This paper presented a numerical simulation technique for calculating the strain and the stress state during one or more sliding cycles at every nodal point of the surfaces containing real roughness topology and layered structure. Depending on the nature of deformation, one can evaluate the actual stress and strain state either from the point of view of the fatigue failure or from the point of view of the exhaustion of deformation capacity of the surfaces.

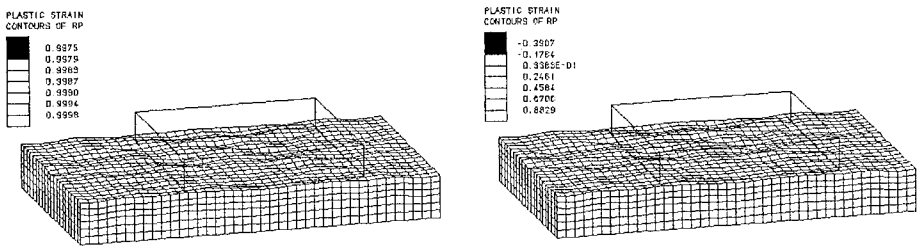


Fig. 11. The values of the reserve of plasticity (RP) before (left) and after (right) the beginning of the particle detachment by ductile fracture

The presented calculation results show that the contact remains elastic until the real contact area is sufficiently large to distribute the contact pressure. Differently from this, the plastic deformation of the roughness peaks begins and – based on the presented results – the plastic strain, considering the strain hardening process, exceeds the deformation capacity of the base material even in the case of the first plastic contact. Using the flow power strain hardening law of the basic material and limiting the maximum plastic strain by the deformation capacity, which was calculated from the specific fracture work, it was not possible to demonstrate plastic ratchetting. To simulate the ratchetting mechanism, it is necessary to ensure that the surface be able to tolerate large plastic strains to fracture, i.e. one has to define perfectly-plastic properties compared to the substrate for the surface layer. The perfect plasticity may be a consequence of the high hydrostatic component of stress or of the superimposed hydrostatic pressure in concentrated contacts, but this phenomenon could not be verified by the presented model under dry friction conditions.

Our numerical model is not able yet to simulate quantitatively the wear, but it is possible to predict the probable degradation mechanism of the surfaces considering their nominal load and the required lifetime. Through this methodology, it would be possible to optimize the manufacturing of the surfaces or the use of antiwear coatings in the future.

Acknowledgement

We are grateful to Prof. J.A. Williams from the University of Cambridge for reviewing our manuscript and for his remarks.

This work was supported by the Hungarian Scientific Research Fund (OTKA) under project No. T 016007.

Nomenclature

$\sigma_{x\ \mathrm{res}}$ – residual (initial) stress
E – Young's Modulus
ν – Poisson's Ratio
$\overline{\sigma}$ – flow stress
$\overline{\varepsilon}^{pl}$ – equivalent plastic strain
σ_0 – yield stress
σ_D – fatigue limit
σ^* – maximal tensile stress
$\overline{\varepsilon}_f$ – deformation capacity
W – specific fracture work
W_e – work of elastic deformation
W_p – work of plastic deformation
W_c – work of contraction
ε_m – limit of uniaxial (linear) strain
σ_B – tensile strength
σ_m' – actual stress at maximal tensile load
Ψ_c – contraction at fracture
Ψ_m – contraction at the end of the linear strain
σ_m – actual (true) stress at the end of the uniaxial strain
F_m – maximum load
d_0 – initial diameter of the tensile specimen
d_1 – diameter of the tensile specimen at the end of the uniaxial strain
d_m – diameter of the cylindrical part of the tensile specimen at fracture
ε_c – contraction strain
α – factor describing the deformation history ($= \mathrm{d}\varepsilon_2/\mathrm{d}\varepsilon_1$)
K – constant of the flow power strain hardening law
n – exponent of the flow power strain hardening law
RP – reserve of plasticity

References

[1] GODET, M.: Third-Bodies in Tribology, *Wear*, **136** (1990), pp. 29–45.
[2] BERTHIER, Y.: The Third Body Concept: Interpretation of Tribological Phenomena, *Tribology Series, 31, Proceedings of the 22nd Leeds-Lyon Symposium on Tribology*, Elsevier, (1996), pp. 21–29.
[3] SEMENYUK, N. F. – BACHINSKAYA, N. K.: Analytical Model of Rough Surfaces Contact, *Journal of Friction and Wear*, **14**, No. 5, (1993), pp. 24–28.
[4] KOPALINSKY, E. M. – OXLEY, P. L. B.: Explaining the Mechanics of Metallic Sliding Friction and Wear in Terms of Slipline Field Models of Asperity Deformation, *Wear*, **190** (2), (1996), pp. 145–154.
[5] TORRANCE, A. A. – BUCKLEY, T. R.: A Slip-line Model of Abrasive Wear, *Wear*, **196** (1–2), (1996), pp. 35–45.

[6] KAPOOR, A. – JOHNSON, K. L. – WILLIAMS, J. A.: A Model for the Mild Ratchetting Wear of Metals, *Wear*, **200** (1996), pp. 38–44.
[7] VÁRADI, K. – NÉDER, Z. – FRIEDRICH, K.: Evaluation of the Real Contact Areas, Pressure Distributions and Contact Temperatures During Sliding Contact Between Real Metal Surfaces, *Wear*, **200** (1–2), (1996), pp. 55–62.
[8] STALIN-MULLER, N. – DANG VAN, K.: Numerical Simulation of the Sliding Wear Test in Relation to Material Properties, *Wear*, **203–204** (1–2), pp. 180–186.
[9] NOGI, T. – KATO, T.: Influence of a Hard Surface Layer on the Limit of Elastic Contact – Part I: Analysis Using a Real Surface Model, *Journal of Tribology*, **119** (1997), pp. 493–500.
[10] XIE, Y. – WILLIAMS, J. A.: The Generation of Worn Surfaces by Repeated Interaction of Parallel Grooves, *Wear*, **162–164** (1993), pp. 864–872.
[11] CHIZHIK, S. A. – GORBUNOV, V. V. – MYSKHIN, N. K.: Computer Modelling of the Solids Contact Zone Based on the Scanning Probe Microscopy, *Journal of Friction and Wear*, **14**, No. 4, (1993), pp. 5–12.
[12] ANDO, Y. – INO, J.: The Effect of Asperity Array Geometry on Friction and Pull-Off Force, *Journal of Tribology*, **119** (1997), pp. 781–785.
[13] LUNDE, L. – TONDER, K.: Numerical Simulation of the Effects of Three-Dimensional Roughness of Hydrodynamic Lubrication: Correlation Coefficients. *Journal of Tribology*, **119** (1997), pp. 315–317.
[14] LUNDE, L. – TONDER, K.: Pressure and Shear Flow in a Rough Hydrodynamic Bearing, Flow Factor Calculation. *Journal of Tribology*, **119** (1997), pp. 549–555.
[15] AI, X.: Effect of Three-Dimensional Random Surface Roughness on Fatigue Life of a Lubricated Contact, *Journal of Tribology*, **120** (1998), pp. 159–164.
[16] SZABADÍTS, Ö., – KRÁLICS, GY. – LOVAS, J.: Bestimmung zyklischer Fließkurven zur Berechnung von Blechumformvorgängen, *Blech Rohre Profile* **41**, 3 (1994), pp. 190–194.
[17] YU, C. C. – KEER, L. M. – STEELE, R. K.: Three-Dimensional Residual Stress Effects on the Fatigue Crack Initiation in Rails, *Journal of Tribology*, **119** (1997), pp. 660–666.
[18] KRAGELSKI, I. V.: Reibung und Verschleiß, VEB Verlag Technik, Berlin, 1971.
[19] ZHARIN, A. L. – SHIPITSA, N. A. – FISHBEIN, E. I.: Some Features of Fatigue at Sliding Friction, *Journal of Friction and Wear*, **14**, No. 4 (1993), pp. 13–22.
[20] KAPOOR, A.: A Re-evaluation of the Life to Rupture of Ductile Metals by Cyclic Plastic Strain, 1994, *Fatigue Fract. Engng. Mater. Struct.* **17**, No. 2, pp. 201–219.
[21] JOHNSON, K. L.: Contact Mechanics and the Wear of Metals, *Wear* **190** (1995), pp. 162–170.
[22] FLEISCHER, G. – GRÖGER, H. – THUM, H.: Verschleiß und Zuverlässigkeit, VEB Verlag Technik, Berlin, 1980, pp. 124.
[23] ZIAJA, GY. – CSER, L. – DARVAS, Z.: Simulation and Limit State in Metal Forming, Advanced Technology of Plasticity, *Proceedings of the Second Conference on Technology of Plasticity*, Ed. Lange, K., Springer Verlag, Berlin etc., Vol.II., (1987), pp. 999–1003.

PERIODICA POLYTECHNICA SER. TRANSP. ENG. VOL. 28, NO. 1–2, PP. 91–102 (2000)

BUDAPEST – ON THE WAY TO JOIN THE IDIOMA PROJECT

Katalin TÁNCZOS and Péter RÓNAI

Department of Transport Economics
Faculty of Transportation Engineering
Budapest University of Technology and Economics
H–1521 Budapest, Hungary
Tel.: 36 1 463 1008; Fax.: 36 1 463 3267
e-mail: ktanczos,pronai@kgazd.bme.hu

Received: Oct. 5, 2000

Abstract

As traffic volume increases continuously, problems appearing in this field need new kind of solutions. Old freight operation methods are often not satisfactory to the needs of new demands. This paper will show the opportunity for the capital Budapest to join the EU project in the fields of city-logistics: "Innovative Distribution with Intermodal Freight Operation in Metropolitan Areas (IDIOMA)". Summarising the objectives and the general project scope will help to find connection points to the current shape of freight transport in Budapest. Providing, analysing important data, and specifying the stressed fields for logistic operations the paper finishes with the already reached steps of the project assessment contribution.

Keywords: freight centres, intermodal transport, city logistics, waste management.

1. Introduction

The actual high transport volumes will still increase in the next years - some prognostics assume a growth of 100Union. In particular, in urban centres the volume of traffic is causing already today serious environmental damage and health problems. The unbearable level of traffic worsens living conditions and, consequently, the quality of life of many citizens. As mobility of goods and persons is a base for economic wealth, innovative solutions are to be found to cope with the increasing traffic. Cities and regional administrations are forced to think of measures to support alternative transport concepts and to implement policies which will reduce the pressure on the city infrastructure and the inhabitants caused by the traffic.

2. The Objectives and Scope of the Project

IDIOMA (Innovative Distribution with Intermodal Freight Operation in Metropolitan Areas) is a demonstration project funded by the European Union in the Fourth Framework Programme contributing to find solutions for the increasing goods traffic in metropolitan areas. It is running from December 1998 to February 2001. With its five demonstration sites – Nuremberg (Germany), Öresund (Sweden/Denmark),

Paris/lle de France (France), Randstad (Netherlands) and Zurich (Switzerland) IDIOMA covers a wide range of projects. All demonstration sites have in common the co-operative approach of the projects: Both forwarders operating on road and/or rail and public administrations contribute to reduce the negative effects of goods traffic in urban areas [1].

The general objective of IDIOMA is to demonstrate the options to improve the distribution of goods within metropolitan areas and between intermodal transport terminals/freight centres and metropolitan areas. In particular this means that the demonstration tasks will show solutions for:

- the more efficient operation of freight centres by optimising the information flows to ease the transhipment of freight between the available modes and means of transport and the use of multimodal chains;
- the integration of small and medium sized enterprises (SMEs) and their specific requirements in freight centres and transhipment point concepts to push their active participation in offering intermodal transport services and
- the optimum organisation and the integration of telematics in the distribution process before or after a long haulage intermodal transport leg to achieve sustainable freight flows in urban areas and to reduce the environmental impact.

IDIOMA subsumes different situations at five European test sites with different prevailing conditions. It supports the exchange of experience between the participating actors and encourages other cities to benefit from its results and to implement measures for a more efficient and environmentally sound freight distribution.

The IDIOMA project is realized in commission by DG VII (its new name is DG TREN: Transport and Energy) and DG XIII (Telematics) of the European Commission.

The IDIOMA project aims at influencing the policy process by showing the best practices in the area of intermodal urban freight transport. The objective of this chapter is therefore to present a general overview of European policies in the field of urban freight transport including intermodal transport in cities. The initial problem is that there is not one single urban freight transport policy in the EC to describe but many urban freight transport policies. Urban freight transport is more than other transport activities, it is a subject of local, regional and national policies in different policy fields, such as transport planning, environmental planning and economic planning. The policies, therefore, are different in the various countries and have changed over the time.

A few figures are to be mentioned about urban Europe. Around 20population live in large conurbation of more than 250,000 inhabitants, a further 20in medium-sized cities of 50,000 to 250,000 inhabitants, and 40areas with 10,000 to 50,000 people. Important differences in economic structure and functions, social composition, population size and demographic structure as well as geographical location shape the challenges which urban areas face. National differences in traditions and culture, economic performance, legal and institutional arrangements and

public policy have an important impact upon cities and towns. There is no single model of a European city.

3. Budapest, as a Follower City of the Project

To the follower city activities belong mainly the following four objectives:

- to make an overview of the status of the city of Budapest in the field of urban goods transport and pre- and endleg of intermodal activities (presented in Chapter 4);
- to present IDIOMA developed ideas in the context of Budapest;
- to present experience and ideas from Budapest on the pre- and endleg of intermodal activities;
- to provide results to the final IDIOMA reports.

In order to achieve the mentioned aims, the following steps have to be worked out:

- To overview the status of the city, Budapest with special respect to its freight flows and freight traffic distribution system (already done,to be demonstrated in the following chapter).
- To identity the actors, the possible stakeholders of the system (already done, as shown in the following chapter).
- To work out a questionnaire to be distributed and completed by the potential participants.
- To collect data and produce them.
- To compose and analyse different clusters of the freight traffic flows according to the type, the place, the quantity, the quality and the frequency of the transported goods.
- To identify the bottlenecks of the freight traffic flows considering the capacity of the transport network of the city.
- To make proposals for increasing the effectiveness of the transport chains, the implementation and evaluation of the economic efficiency. The dissemination of the results.

4. Overview of the Available Research Results

4.1. Transport Policies, Changes and Main Processes

The transport development of Budapest is an organic part of the compact urban development process which aim is to increase the quality of life in parallel to the operational efficiency of the city, with a careful consideration of the conservation of the natural and built environment.

The main principles of the existing development plan of the city are:

- unification of regional and transport planning,
- integrated network and operation planning,
- managing the demand for mobility,
- regulation of the direction and intensity of the traffic flows,
- user-friendly installation of the transport infrastructure,
- harmonisation of the quality of transport and environment [2].

The new approach of the development plan combines searching for the best solutions for construction of the missing elements of transport infrastructure with more efficient utilisation of the existing ones and continuous management of the demand for mobility.

Due to the geopolitical position of the capital is attributed an important role to the transport system. The development of Budapest was in the past and will be in the future always determined by the quality of its transport connections. Budapest may become an international centre of the region because of its excellent multimodal links to other main and subcentres of larger distance, like its road and rail connections, supported by the waterway of Danube and the Ferihegy international airport (see the IVth, Vth, VIIth and Xth transport corridors of the Trans European Network). The transport network of the region is suitable for creating the conditions of a production, consumption and distribution center.

The economy of Budapest is continuously developing and changing. 35GDP of the country has been produced here. The number of operating companies in the capital is more than 230 thousand. 60billion USD) in the last decade was concentrated here. The changing structure of the economy and employment has reshaped the layout of the region and the characteristics of the mobility. The city inherited a rather well developed road and public transport network.

The political and economic changes in the last decade had a significant influence on the transport system of the city. In the previous regime, public transport played a dominant role in the urban living conditions. Public transport provided the mobility for the majority of the population with very good accessibility, frequency and low level of tariff.

After the radical changes in the economic and political systems, the modal split became significantly worse. According to the last household survey the ratio of the public transport and the individual transport is 60 to 40% while it was 80 to 20% at the end of the last decade.

The major problems are the followings:

- car ownership has grown significantly,
- more intensive use of cars in some entrepreneur groups,
- formerly state-owned large-size companies have been closed and thousands of SMS new companies have been established,
- new activities appeared in the city centre (offices, banks, insurance companies),
- more demand for than supply parking facilities.

Changes taking place in the economy and commerce had a multi-part effect to the good-supply. As production and good transport decentralised, the number of small enterprises increased, and this caused the decentralisation of good supply operations. At the same time, some shopping centres have been built, particularly in the outskirts of the capital, which means a kind of concentration of supply operations. All together, today is represented by the continuous changing position of commerce and goods transport, having now an intermediate shape: beside the historical old methods there are examples of ordering systems using virtual communication and the world-wide web.

In the capital there is no well organised supply method, collecting and distributing system and the goods supply for shops is different in nearly all kinds of articles. The increased number of goods supply transports causes bigger and bigger traffic on the roads. The most characteristic feature of freight transport is the lack of organisation, the usual tool for regulation is making restrictions. The greatest part of commercial institutions is still in the city, this is why the traffic increasing effect of good transport, loading and unloading causes the greatest problems in the most jammed part of the capital.

Shopping centres in the outskirts have a decreasing effect to the transports directed to the city, but this is balanced by other processes, so the volume of supply traffic had only a small change during the past years. The number of the drive in permissions handed out in the city has showed no change at all. The reasons for this are:

- restructuring shop profiles to reach more customers, and
- possible decline of some goods' proportional rate in several shops has no influence on the number of transportation vehicles.

Still, shopping centres accommodated in the outskirts attract significant freight traffic with an average of 3,000-3,400 vehicles per month, which means about 100-140 vehicles a day. 54% of the whole transport needs is carried out by freight vehicles, the second biggest part belongs to smaller freight vans (42%), and only 4% to trucks.

As a result of the increasing number of shopping centres, the last part of the supply chain is carried out by customers and inhabitants, taking the goods from the shopping centre to home by own car. This part of freight transport cannot be monitored by the ordinary methods of logistics administration, at the same time it provides a significant surplus in road traffic.

4.2. Important figures in general and about the freight traffic in Budapest

At the end of the 20th century 62% of the total population of Hungary live in cities and towns. The distribution of the population in the cities of different size reflects the specific impacts of the history of the country. More than one fifth of the total population, 2 million inhabitants live in the capital and in its agglomeration but the second largest city of the country, Debrecen, has only 200,000 inhabitants and

besides these two ones, only seven other towns of the country have more than one hundred thousand inhabitants.

Data from 1997 shows that food consumption in Budapest was 700 kg/person/year which means about 2 kg/person/day. To cover this demand, in the whole year the city reached 128,000 tons of goods as total freight traffic performance (this makes about 430 tons per day). For the whole area of Budapest this number is 1,303,000 tons per year (4350 tons per day). Waste taken out from the capital was 4,450,000 m3 per year (1997) of which municipal waste was 2,880,000 m3 (about 5 litres per person per day).

30% of all loading and unloading processes in Budapest take place on the public roads, 67% of these processes using dedicated loading and unloading places. The traffic disturbing effect of freight transport and loading is increased by the fact that 90% of all freight transport is carried out during working hours. Receiving goods not during the opening hours reaches not more than 10 percent of all freight transport. In the town exist 421 dedicated loading and unloading places, each of which belongs to 3-4 shops or store.

The whole weight transported within the city is 17% higher than the one transported from the country to the capital, at the same time, traffic performance (thanks to the three times longer transport routes) from the country to Budapest is twice higher than the traffic performance within the town. Average transport distance within the capital is 29 km, from the country to the capital the figure shows 77 km. Taking the country data into account, the transport distances within the capital are too long.

During the whole year the percentage of empty running vans does not fluctuate, within the city this figure is 45.5%, coming from the country 33.4% as an average. This means a low average utilisation of vehicles within the capital.

About railway freight transport is important to mention that 60% of all station freight traffic in Budapest is unloaded, the rest (that 40%) is loaded goods. This shows, that the net flow regarded the railway freight transport is coming to the capital. However, railway freight transport does not play a very important role in the town's freight flow: only 14.3% of all incoming weight is transported by trains, which hardens the question of air pollution.

4.3. Simple Solutions

Loading places are signed with suitable boards, but the effectiveness of these boards is poor: parking cars occupy the free place frequently. Because of this, loading is carried out using the traffic lane, disturbing the traffic significantly.

The zone-system introduced in the eighties has the task to regulate the dispersion of freight traffic in time and space, decreasing (and in several critical places even terminating) transit freight transport. Full truck weight restrictions implemented in the zone-system push heavy truck traffic to the low traffic hours in order to establish a possible undisturbed daytime movement on roads.

Since 1996 there is a new full-truck-weight-limit system regulating goods transport:

- in Pest between the upper quayside (felső rakpart) and the middle ring road ((Nagykörút), in Buda between the upper quayside and the Buda Ring Road (Budai körút) there is a 3,5 ton full truck weight limit from 07 a.m. till 06 p.m.
- between the middle ring road (Nagykörút) and the outer ring road (Hungária körút) there is a full-weight limit of 12 tons again from 07 a.m. till 06 p.m.
- in the 9th district in Buda has been introduced a full-weight-limit of 12 tons. The characteristic freight transport routes and the full-weight-limit zone-system is shown on *Fig. 1*.

There are no overall and exact data available in connection with the freight, material and waste transport in Budapest and its agglomeration. As a result of more measurements systematically carried out on the main roads, the volume, dimension and composition of freight traffic are well known. The last direction flow measurement was carried out in 1992-1994, but since then the flow relations and parameters have been changed considerably. To follow these, new complex traffic measurements are taking place in the capital and its region. Data already available is processed permanently. Important data can be found also in the "Traffic Development Plan of Budapest", too.

4.4. Important Freight Transport Centres

Starting and destination points of the freight transport in Budapest are mainly old industrial areas, where reconstruction aims, first of all, at logistic, warehousing and freight traffic centres. These areas are mentioned in town development plans:

- industrial basin Nagytétény (Nagytétényi ipari medence),
- Csepel island,
- logistic area Soroksár,
- Kőbánya, Jászberényi-Maglódi street,
- Cinkotai street-Bökényföldi street,
- Újpalota, Késmárk street,
- Rákospalota, Székely Elek street,
- industrial basin Kaszásdülő (Kaszásdülői ipari medence).

These will provide the centres and sub-centres of city logistics in the future, as sown on *Fig. 2*.

The logistical potential of the industrial basin Nagytétény is based on the Danube as a big-capacity waterway, the railway lines to Székesfehérvár and Pusztaszabolcs, the access part of main road number 6 (later on a motorway), and the southern section of the ring road M0. Apart from these, favourable public traffic connection is available between the logistic area and the city centre. First of all, the Budapest Harbour Park is going to be built out on enterprise basis, after this

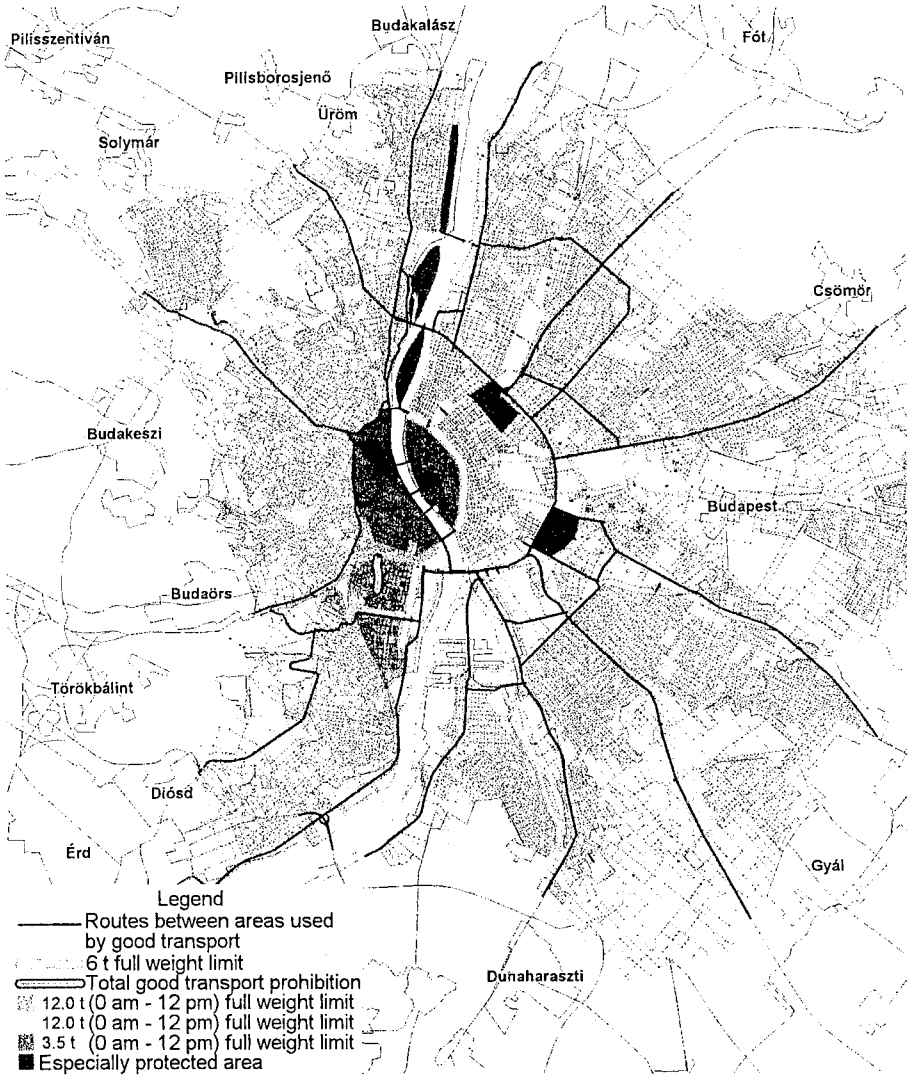


Fig. 1. The different freight zones in the capital

the ROLA and RORO terminal is going to be established, working as a "logistic engine" of the area. On the field of the former pig farm in Nagytétény logistics centre will be built out in two steps, finally 250,000m2 as total covered surface. The container- harbour planned next to the venture park will be established in the first step on a field of 10 hectares. Co-operation and function sharing will take place between the two area. The two project can help to change function of the empty areas of the industrial basin Nagytétény.

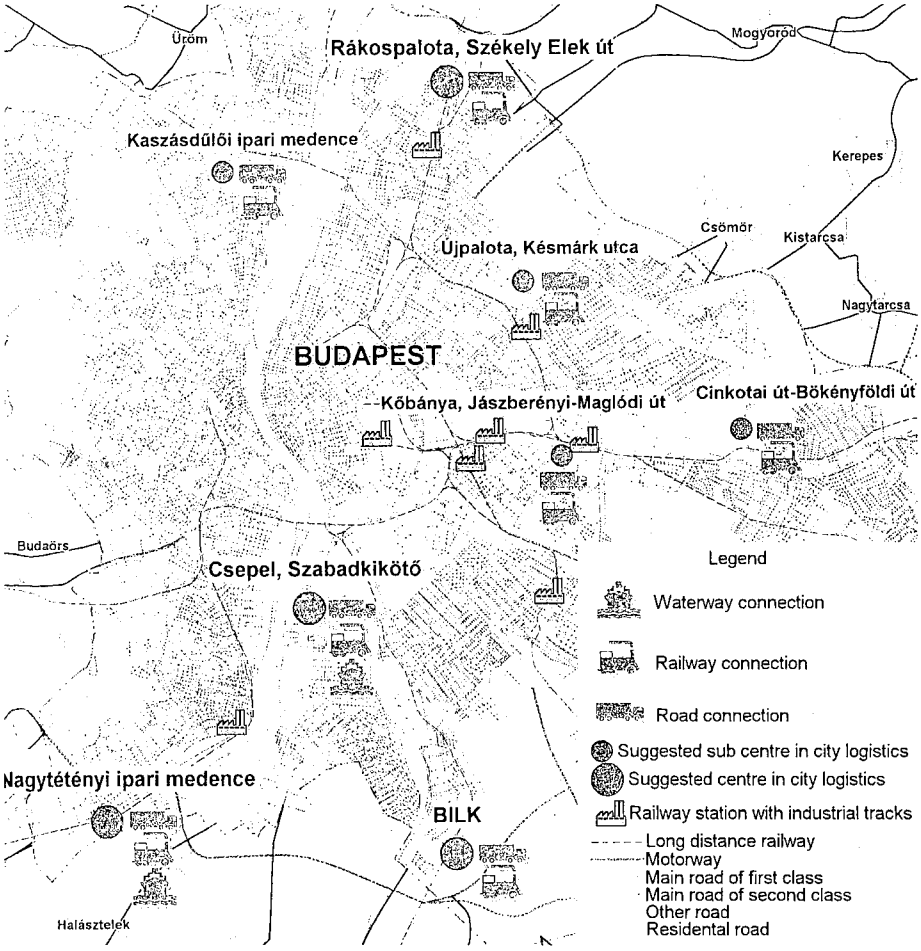


Fig. 2. Planned important logistic junctions in Budapest

The most important facility of the Csepel area is the Danube harbour. Right now it is only possible to access it trough a built up area. Rail traffic also disturbs the inhabitant traffic in the district even if rail traffic volume has been decreasing for nearly 10 years. Road connections between the area and the city are acceptable, but industrial and public transport between this region and other industrial/logistical areas have to be developed soon. If freight traffic increases, cross-Danube road capacities might need further developments. The harbour will have to act as the waterway connection for the Budapest Intermodal Logistic Centre (BILC) in Soroksár.

Heart of the Soroksár logistic area is the already mentioned BILC, some parts of which are already realised (railway connection tracks), and planning is in the final

stage. The area can be found between the southern section of the ring road M0 and the railway line to southern Hungary (Kelebia). Road connection with the city centre is well established. The BILC is going to be the most important logistics centre of Hungary, and will have tasks coming from other central European countries, too. Areas around BILC are now changing their function, approaching auxiliary logistical duties. Already more logistic service providers as well as shipping agencies, and warehouses have settled down in the region over the old industrial fields. As the district is in possession of free manpower, besides the central BILK project more other entertainment companics are expected to settle down in the neighbourhood.

Connections between northern Kőbánya and the city or other freight traffic sources are not suitable. As the eastern section of M0 does not exist yet, the yard can only be served trough residential areas. In harmony with earlier ideas about the line of M0, next to the Tündérkert street warehousing capacities have been created. The area still has great reserves, much of which will be only used after improving road connections to other parts of the town. After saturation, free capacity can be offered by the northern industrial area of the 10th district (Jászberényi street – Maglódi street). Investments on this ground, however, need higher capital for pulling down old industrial plants.

The Cinkotai street – Bökényföldi street has even worse traffic connection with other parts of the capital than the previously mentioned Kőbánya. Still, in the district settled environment already at its starting time contained a wide range of important transport companies (like Hungarocamion, Budapest Traffic Company, Hungarian Post). Important logistical role can be realised only after building out the eastern section of M0.

Késmárk street in the 15th district has an advantageous position, but is to be found between residential areas, which has reduced the expansion options of incoming companies. Although the starting section of the M3 motorway is near the yard, it can only be reached trough residential areas, disturbing inhabitants.

Among industrial grounds in the northern – eastern part of Budapest, Rákospalota has the most favourable features. The warehouse district lies between the M0 ring road and the railway line to Szob. Yards next to the Székely Elek street have been utilised by warehousing and wholesale firms. Further enterprises of this kind are expected in the future. In short run, rapid development in the road infrastructure is expected, which will facilitate the "moving in" activity of some new companies. The warehousing basis can be soon reached without disturbing residential areas.

Basin Kaszásdülő is "closed" from all sides, it is hard to reach. After building out the M0 ring road it will be possible to separate industrial traffic, but the connection to other industrial areas will be still behind the desired standard. Changing the old industrial functions, some new enterprises have already moved in, providing logistic services, warehousing and wholesale trade.

The introduced areas are developing differently, but all of them show a significant movement towards logistics. In many cases relationship to residential areas and environment determines the chances of companies settling down here. The classic logistic services exist only in the area of southern Budapest, where ordinary connections with the city and other areas are already established. Firms using the

northern and eastern areas mainly are wholesale traders, only with poor logistic services [3].

4.5. Data about waste management in the capital

Systematic waste gathering takes place for nearly all flats in the capital (99,8Q%). Since 1990 the transported volume of waste increases by 10% every year, in 1998 it was 4,360,000 m^3 per year. The task is performed by 150 special waste carrying vehicles, which have a load weight of 24-26 tons and have 3 axles. Daily transport performance is about 85 m3 per car. In average, the vehicles are 10 years old, but their technical condition meet the European standards. 60% of the waste arrives in the waste-burning plant in the 15th district, 40% in the waste depot between Dunakeszi and Fót. Average transport distance is about 21-25 km.

Collecting waste is carried out with 50 vehicles in the city between the middle ring (Nagykörút) and the Danube from 5 till 7 o'clock in the morning, after this it is continued on the Buda side and between the outer ring (Hungária körút) and the middle ring (Nagykörút). Vehicles move after a planned timetable, full loaded vehicles reach the burn-plant using an advised (but not compulsory) route. Traffic jams happen often in the morning in the city and in one way streets outside the middle ring. Many times happens that waste loading needs the traffic lane, too, this makes difficulties for drivers. Data about systematically emerging traffic jams have not been collected yet [4].

4.6. Stakeholders of the project Budapest

Main actors of the project has to be identified in order to be able to organise work. Different parties have to be represented among the stakeholders, otherwise it is not possible to collect all necessary data, to analyse them, and to find innovative solutions for the key problems. Following scientific, governmental institutes, and private companies have been considered as interested participants in the project:

- The project of "Follower City Budapest" will be lead by the Department of Transport Economics at the Budapest University of Technology and Economics. The Department of Transport Operation of the same university will be also an active contributor to the project.
- The Logistics Directorate of the Ministry for Economics will be participating in the project as the representative directly of the government and indirectly of the legal regulation.
- The Transport Directorate of the Municipality of Budapest responsible for managing transport related issues of the capital and its agglomeration.
- The freight transport operating companies of the region (e.g. Hungarian Railways, etc.).
- ESRI Hungary as the representative of the IT companies.
- All the private commercial and retailer companies working in the region.

5. Summary

As shown in this short report, the Hungarian capital, Budapest has many tasks, problems, difficulties about freight management, but, at the same time several solutions are already provided, and at some places significant developments have been undertaken. To play the role of a great metropolis many questions have to be answered in the field of supply chain management, logistic centres, service providers, private enterprises and town traffic management. The town government has to play an important role, but many other tasks can be performed by private companies. Developing adequate methods in co-operation with the existing human resourses and education can help to obtain a more effective social yield provided by the recovery of the Hungarian economy.

Already reached project results are:

- detailed description of current freight transport volumes,
- main freight transport routes within the capital Budapest,
- possible logistic and goods transport centres and their circumstances,
- identified stakeholders.

As shown in chapter 3, there are 5 more steps to be fulfilled till the project finishes. The scientific results can be the topic of the next published paper. These should show solutions for:

- a more efficient operation of freight centres by optimising the information flows, to ease the transhipment of freight between the available modes and means of transport and the use of multi-modal chains,
- the integration of SMEs and their specific requirements in freight centres and transhipment point concepts, the offer intermodal transport services and,
- the optimum organisation and integration of telematics in the distribution process before or after a long haulage intermodal transport leg to achieve sustainable freight flows.

References

[1] IDIOMA, Design process results, project paper, PTV Plannung Tansport Verkehr AG (as contractor), 2000.

[2] Katalin Tánczos, D.Sc.: Applying research in defining the infrastructure needs, conference paper, CARISMA – Transport Central and Eastern European Forum, Budapest, 2000.

[3] [3] A városi áruszállítás logisztikai megszervezésének lehetõsége Budapesten (Possibilities for organising logistical goods transport in Budapest), Közlekedés Fõvárosi Tervezõ Iroda Kft (Capital Transport Planning Bureau Ltd., study in Hungarian) 1999.

[4] Budapest városfejlesztési koncepciója (Budapest Town Development Plan, Városkutatás Ltd., (study in Hungarian) November 1999.

PERIODICA POLYTECHNICA SER. TRANSP. ENG. VOL. 28, NO. 1–2, PP. 103–116 (2000)

PROBLEMS AND ISSUES OF DEFINING USER CHARGES FOR RAILWAY INFRASTRUCTURE

Katalin TÁNCZOS and Gyula FARKAS

Department of Transport Economics
Budapest University of Technology and Economics
H–1521 Budapest, Hungary
Fax: +36 1 463-326
Phone: + 36 1 463-1008
e-mail: farkas@sc.bme.hu

Received: May 24, 2000; Revised: September 22, 2000

Abstract

The EU Directives accepted in the 90s determined the opening of the Rail Market and the constitutional separation of the European railways (railway operator, *infrastructure manager*).

This article specifies the methodological issues of the Hungarian charge system, demonstrates the European and national practice of the determination of the user charges for railway infrastructure and intends to show the tasks, which have to be done in the near future.

Keywords: user charges for railway infrastructure, infrastructure manager, service standard.

1. Introduction

In the late 80s it became unambiguous that transport (operating of the transport systems) had to be placed on a new ground. Therefore operation improvement and modernisation of the railway transport systems – on the basis of the same principles for every member state – were started.

The 91/440 EU Directive required separating of accountancy, in some countries the effective taking apart (infrastructure; operator) was realized – in Sweden, United Kingdom – or arrangements were made towards it – Germany. [1] Some countries revised their own railway operator and prepared different strategies for the improvement of efficiency (privatisation, cost and staff reduction). In connection with charges for infrastructure the following participants, problems and issues had arisen (*Table 1*).

This article demonstrates and analyses the German and Polish practice of user charges for railway infrastructure and after examination makes a proposal concerning the methodological issues of the Hungarian charge system.

Table 1. Participants and issues in relation with railway infrastructure

Participants	Issues
Infrastructure owners, regulators	Who is paying?
Infrastructure manager	Who should be paying?
User	How much should be paid?
Third partner	How? On what grounds?
Society	What should be paid for?

2. Practice in Germany

Traditionally, the transport sector in Germany was strongly influenced by the government. It usually led to disputes on the market of services. However, for treating it, especially the reasons were always well founded: public welfare, public utility services, externalities and monopolies. As the academic discussions and political agreements have accepted the need of deregulation and privatisation, the first steps towards execution were made [2].

'Private-public-partnership' supported projects – which had been previously used in transport and post office – appeared in the railway system.

After the East-German Railway had merged into the German Railway (DB) they opened their network for a third partner (1999). The first year of introducing their user charges for railway infrastructure was 1994.

The previously mentioned charges in the German Railway Transport had been surveyed and analysed in another study – where the main features of the competition had been discussed.

The most important steps in the reform of the German Railway were:

- separation of the infrastructure and the operation;
- opening of the network for third partners by introducing the user charges for railway infrastructure;
- regionalization of the suburban transport.

Table 2 shows the role and the structure of the different railway transport organizations in 1995.

The German Railway Ltd (DB AG) and other railway companies offered usage of the railway public service in return for user charges for railway infrastructure. The DB AG has made usage of railway network possible for carriage of goods and travel agencies too. Access to the railway infrastructure for different users has been provided in the most comprehensive way in Germany. While there is a lack of regulation framework for a long time in the majority of the member countries of EU, the German regulation on the usage of the railway infrastructure was put into practice very early. The new regulation includes the following:

Table 2. The role and the structure of the different railway transport companies in Germany 1995

	DB AG	Regional railway company	Sum total
Number of companies	1	99	100
Length of network (1000 km)	41.7	3.4	45.1
(percentage)	90	10	100
Number of employees (1000)	312	16	328
(%)	95	5	100
Passenger traffic (million passengers)	1330	323	1652
(%)	80	20	100
Passenger traffic (million pass kms)	60400	3110	63511
(%)	95	5	100
Freight traffic (million tons)	296	56	352
(%)	83	17	100
Freight traffic (million ton kms)	67910	880	68790
(%)	88	12	100
Returns (million DM)	26910	1540	28450
(%)	95	5	100

- every railway company which provides access to the railway infrastructure can determine the level and the structure of the fee for third partners;
- there is no independent authority which can regulate charges;
- they can settle average charges for the whole or some part of the network as well as for specified sections;
- the following factors can be taken into account when charges are determined:
  - type of running vehicle,
  - usage of lines,
  - usage of capacity of lines,
  - noise and air pollution caused by trains.
- the infrastructure charging system has to be demonstrated for the customers but it is not necessary to be published;
- customers will be rewarded by a price reduction offered by different railway companies, if their order for train km performance is above a certain level;
- if particular customers compete with each other for access to the infrastructure it is allowed that the best offer will be accepted.

Parallel to the introduction of user charges the railway companies were obliged to cover all the expenses including network operation and maintenance as well as depreciation. *Table 3* shows that the originally introduced user charges for railway infrastructure consisted of two sections: partly basic price, which was corrected

by customer requirements (reliability, special types of train), partly by reduction factors depending on the used total train km performance during the period of contract. The total user charges for infrastructure included the user charges for station, too, calculated on the basis of number of train stops and types of train [3].

Table 3. The structure of the user charges

Base charges

- Route categories

  Importance: 3 categories
  Quality/implement: 7 categories

- Train categories

  Passenger: 7 categories
  Freight: 5 categories

Factor

- Depends on the required time scheduling (0.8 . . . 1.2)
- Depends on the use of the railway track (caused by special train type: 0.9 . . . 1.1)

Reduction depending on order-volume

Use of number of tracks:

- 1% . . . 5% local passenger trains
- 1% . . . 20% long-distance passenger trains
- 1% . . . 20% freight trains
- 2% . . . 10% more than 2 years contractual commitment to the ordered infrastructure

The main problem in the previously mentioned version was that the charges were too high, the structure of charging was cross-financing and the corrected coefficients and finally the charges did not cover all the expenses.

Based on the experience the charging system has been revised. The advantage of the new, two-part charging system is that it mirrors the fair cost-structure of the railway infrastructure operation (fixed price components are related to the rail-network fixed-cost loading and independent of the actual use (InfraCard); variable price components allow sufficient freedom for control and flexibility required and are depending on the actual use) and it will assist the aim of the rail reform of attracting more traffic.

Railway operators which enter a long-term – above 2 years – contractual commitment with the infrastructure manager are rewarded with a price reduction (between 2–10%).

Railway operators which have purchased an InfraCard pay a variable price per train path for the corresponding lines (DM/train km). This price is corrected by factors depending on capacity utilisation of the line, timetable flexibility, surcharges and reductions. [4]

The price for the use of low-traffic lines is lower and the price for the bottlenecks or busy lines is higher. *Table 4* shows the value of the factor depending on traffic density classes.

Table 4. Price correcting factor values depending on traffic density

	Traffic density class	Factor
B S	High frequented city lines	1.35
B I	High frequented lines	1.15
B II	Medium frequented lines	1.00
B III	Weakly frequented lines	0.85

An important feature of the modified infrastructure charging system is the consideration of the flexibility for scheduling, granted by the railway operators. More flexibility increases the chances of loading the system economically. Basic interval timetable leads to a very high cost and restricted availability of the infrastructure. More flexible demands for paths cause price reduction. Preconstructed paths offered by the infrastructure manager are the cheapest.

Charge for infrastructure depends on technical standards and traffic density of track. *Table 5* shows the structure of this system.

Table 5. The structure of the charging system

Track categories		Traffic density class		
		High	Medium	Low
K1	Tracks for speeds over 160 km/h	Most expensive		
K2	Tracks for speeds from 120 to 160 km/h			
K3	Tracks for speeds from 120 to 160 km/h*			
K4	Tracks for speeds from 80 to 120 km/h			
K5	Tracks with signalling for speeds up to 80 km/h			
K6	Tracks without signalling for speeds up to 60 km/h			Cheapest

*Mainly for passenger traffic

The system of the additions and reductions will be specified in the near future.

3. Practice in Poland

The basis of users charging for railway infrastructure in Poland has been based on the Minister of Transportation and Maritime Economy decree of 1998. The total route charge consists of several section charges as infrastructure costs change along the line.

The infrastructure expenditure database, which is corrected by the inflation in the previous year, serves as a basis of all calculations.

The infrastructure costs concern maintenance costs, operating costs and administration costs whereas depreciation and investments have not been included. Charge for the access to lines/sections consists of the base charge and the energy charge. Charges for additional services are also included if they have been resorted [5].

The base charge covers the value of basic and mandatory services. Basic services include:

- access right to infrastructure;
- timetable construction;
- necessity of usage of all infrastructure elements;
- power supply;
- traffic management, traffic control and train communication.

Mandatory services comprise:

- traffic safety and security;
- rescue in the case of an accident.

The base charge is calculated by the following formula:

$$\boxed{K^p_{i\ \mathrm{base}} = \beta_{1999/1998} * L_i * k_z * N^p_i * s^p_i * \alpha^p_{i1} * \alpha^p_{i2} * \alpha^p_{i3} * \alpha^p_{i4} * \alpha^p_{i5} * \alpha^p_{i6},}$$

where $K^p_{i\ \mathrm{base}}$: base charge for N^p_i route given to operator to run p train on i line,

L_i : length of line i,

$\beta_{1999/1998}$: ratio of materials, energy and services prices in current year related to former year prices,

k_z : profit rate (max. 1.05; for non PKP[1] operators only),

N^p_i : number of routes thrown open for p train on i line,

s^p_i : unit base charge for p train on section of i line,

$\alpha^p_{i1} - \alpha^p_{i6}$: corrective coefficient defined for p train on i line.

[1]Polish Railway Company

The unit base charge is calculated for every section of railway line as well as every train category. The calculation depends on the infrastructure expenditure borne and the transportation work performed by a given train sort.

In addition the unit base charge is influenced by six corrective factors (technical standard of line, traffic density, hour within the day, week day, punctuality, train weight).

The energy cost consists of energy unit price, energy volume used for traction matters and energy distribution, transfer and transformation components. The energy charge can be defined in two ways. Firstly according to the vehicle's onboard energy meter indications, secondly by separate contract conditions. Using an onboard consumption meter the energy consumption charge is defined on the basis of three factors all together (energy consumption measured; energy purchase unit price; correcting factors). In the other arrangement (no onboard meter) – according to fixed contracts – the energy charge is calculated by the multiplication of price ratio, section length, train gross weight, average traction energy price, average unit energy consumption as well as using other correcting factors.

The supplementary charges are defined on the basis of individual calculations depending on the usage of different optional and offered services.

These charges can be the following:

- shunting and additional manoeuvres;
- extra trains supervision during train stops;
- public terminals, marshalling and formation yards, train serving and repair tracks access – separate contracts with object administration required;
- access to telecommunication and information networks, additional information;
- demands of other operators.

4. Hungarian Methodological Issues of User Charges for Railway Infrastructure

The Act of 1993 on the railways determines the responsibility of the State – among others – to provide the non-local public railway infrastructure and the non-local mass transport by rail qualified as public service. As the mode to take these responsibilities the Act provides a contract between the State and the national railway company (MÁV).

The fee – according to 91/440 EU Directive and the government decision linked with the planned overall reform of MÁV – must be determined on the basis of technical standard and use of infrastructure [6].

Analysing the general practice in Europe it can be proved that infrastructure managers take the real infrastructure expenses as a basis and calculate base charge on this basis – in compliance with the used service standard. The base charge is corrected with additional services, surcharges and they refund the innocent operators in the case of decrease of standard.

The Hungarian methodology uses the following rudiments:

- Railway track expenditures needed theoretically: technical cost (expenditure), standard (outlay requirements) of the railway track according to EU Directives and arrangements
- Real technical standard of the track: comparing the ratio of the real level to former standards (norms).
- Base charge for railway infrastructure: paid by operators, corrected by different surcharges (region-developmental, business-political, furthering modern environment-friendly transport modes, etc.).

A charge-system must be worked out which is suitable for

- train, category of train;
- route;
- statistical section;
- line, part of network;
- category of line;
- total network

fee determination separately for passenger and freight transport.

On account of the 'Rules to calculate costs of production in railway business' the costs of passenger and freight trains must be separately determined in order to avoid cross-financing between the mentioned two main activities [7]. The controlling approach is a very important factor, too. [8]

Fundamentally, user charges for railway infrastructure can be determined by statistical sections and fees should be summarised at any level, route – separately for passenger and freight transport.

The charges to be paid by the users of railway infrastructure are composed of the following elements [9]:

1. Base charge
2. Additional charge
3. Extra charges
4. Reductions
5. Repayments

4.1. Base Charge

The base charge can be calculated in knowledge of the technical standard of the lines and the corrected expenditure for the infrastructure.

The corrected infrastructure costs include the real costs of the infrastructure (maintenance, depreciation and operational costs), the costs of traffic control, the technologically justified additional maintenance and depreciation, the costs of backward improvement and the quantifiable externalities.

The mentioned cost elements constitute the corrected expenditure normative demand of the infrastructure manager. If this claim is satisfied, then the infrastructure manager will provide the gradually improving infrastructure, even if the technical standard of lines is behind the EU-standards.

Table 6 shows the technical parameters and the factors which determine the quality of the railway infrastructure.

Table 6. Technical parameters and service standard determining factors

Technical parameters of the lines (T_p)	Factors which determine the service standard
Speed of line (km/h)	Linear scale
Axle load (ton)	Linear scale
Electrified line	Yes/No
Number of track	Single track; Double-track
Safety of line (Signalling)	Key interlocking
	Key wedging device
	Mechanical
	Mechanical with light signal
	All-relay interlocking (D55, D70)
	Electronic
Train stopping control	Mechanical with light signal
	All-relay interlocking (D55, D70)
	Electronic
Traffic control	Between stations
	Mechanical block
	Automatic block
Socio-economic value of line	Volume of passenger/freight transport

The AGC-agreement is fundamental for the improvement of the '*A*-category' national lines. The AGC-agreement contains the characteristics of the average European railway infrastructure. *Table 7* demonstrates the AGC-standards.

After comparing the Hungarian railway infrastructure with the AGC-standards the backwardness of the technical standard of the Hungarian railtrack can be determined.

Before that comparison the service-package – which is offered by the rail track – should be analysed and factors of the service standard should be defined.

The passenger transport and the freight transport have different requirements towards the infrastructure manager, therefore certain factors of the service standard

Table 7. AGC-standards for the European railways

	Speed of line	Axle load	Electri-fied line	Number of track	Safety of line	Train stopping control	Traffic control
Stat. Section	160 km/h	22.5 ton	Yes	Double-track	All-relay interlocking	All-relay interlocking	Automatic block

have to be separately determined. *Table 8* lists the factors of parameters, which specify the service standard of the passenger and the freight transport.

Table 8. Factors of parameters, which specify the service standard of the passenger and the freight transport

Factors which determine the service standard	Factors of certain parameters (%) (α_j)	
	Passenger transport	Freight transport
Speed of line	α_{1p}	α_{1f}
Axle load	α_{2p}	α_{2f}
Electrified line	α_{3p}	α_{3f}
Number of track	α_{4p}	α_{4f}
Safety of line	α_{5p}	α_{5f}
Train stopping control	α_{6p}	α_{6f}
Traffic control	α_{7p}	α_{7f}
Socio-economic value of the line	α_{8p}	α_{8f}
Total	100	100

Comparison of the national railway infrastructure (technical standard) with the European network is important in several respects:

1. As-is analyses (region and size of the backwardness);
2. Increasing of competitiveness of the Hungarian railway infrastructure (neighbouring countries, improvement of transit routes);
3. Drafting of the strategic goals (EU-corridors, increasing of the transport capacity and capacity utilization);
4. Access to the railway infrastructure and determination of the user charges for infrastructure;
5. How does service standard influence the change (increase/decrease) of the expenditures for infrastructure?

The AGC-agreement parameters can be considered as a basis of comparison with the standard of railway infrastructure (common service standard).

Condition of the national railway lines has to be assessed by considering the AGC-parameters, and after that the service standard of the statistical sections should be determined.

Table 9 contains 'servicing factors (β_b)' (which are characteristic of the sections) and the scales (which belong to certain parameters).

Table 9. Servicing factors and scales

Speed of line (β_s)

β_s = Section max speed/160 $*$ 100%

Axleload (β_a)

β_a = Section max axle load/22.5 $*$ 100%

Electrified line (β_e)

β_e = 100%, if the line is electrified, otherwise β_e = 0%

Number of track (β_t)

β_t = 100%, if the line is double-track, otherwise β_e = 0%

Safety of line (β_{sa})

Signalling	Key interlocking	Key wedging device	Mechanical	Mechanical with light signal	All-relay interlocking	Electric
Servicing factors	β_{sa1}	β_{sa2}	β_{sa3}	β_{sa4}	β_{sa5}	β_{sa6}

Train stopping control (β_{tsc})

Signalling	Mechanical	Mechanical with light signal	All-relay interlocking	Electric
Servicing factors	β_{tsc1}	β_{tsc2}	β_{tsc3}	β_{tsc4}

Traffic control (β_{ts})

Traffic system	Between stations	Mechanical block	Automatic block
Servicing factors	β_{tsc1}	β_{tsc2}	β_{tsc3}

Socio-economic value of the line (β_{sev})

Socio-economic value of the line	Passenger transport (seat-km)	Freight transport (gross ton weight)
Servicing factors	β_{sev1}	β_{sev2}

Total of the service standard of all statistic sections can be calculated (for the

passenger and freight transport) with the following formula:

$$\boxed{\text{Total service standard } (Tss) = \sum_i (l_i * \sum_j Tp_{ij} * \alpha_{ij} * \beta_{ij}),}$$

where l_i : length of line i,
Tp_{ij} : technical parameter of the line,
α_{ij} : factor of standard (passenger and freight transport separately),
β_{ij} : factor of the parameter (passenger and freight transport separately).

The average service standard of the railway network can be defined with the next formula:

$$\boxed{\text{Average service standard } (Ass) = Tss / \sum_i l_i,}$$

where Tss : total service standard,
l_i : length of line i.

The theoretical base charge can be calculated in knowledge of the average service standard and the corrected expenditure for the infrastructure with the following formula:

$$\boxed{\text{Theoretical base charge} = \text{Corrected expenditure for the infrastructure/Ass}}$$

After definition of the planned performance ratio (on the basis of train km) between passenger and freight transport the specific factors must be formed for the determination of the real base charge (separate own performance of the passenger and goods transport).

The specific base charge can be expressed by means of the theoretical base charge and the total service standard:

$$\boxed{\begin{aligned} &\text{Specific base charge}_{\text{passenger}} = \text{Theoretical base charge}_{\text{pass.}} / Tss_{\text{pass.}};\\ &\text{Specific base charge}_{\text{freight}} = \text{Theoretical base charge}_{\text{freight}} / Tss_{\text{freight}} \end{aligned}}$$

Their specific base charge must be re-multiplied by the real own performance of the passenger and goods transport and after that the base charge (passenger and freight separately) can be defined to the statistic sections, levels, routes and network.

$$\boxed{\begin{aligned} &\text{Base charge}_{\text{section } p,f} = \text{Specific base charge}_{p,f} * l_i * \sum_j Tp_{ij} * \alpha_{ij} * \beta_{ij};\\ &\text{Base charge}_{\text{route } p,f} = \sum_i \text{Specific base charge}_{p,f} * l_i * \sum_j Tp_{ij} * \alpha_{ij} * \beta_{ij} \end{aligned}}$$

The base charge includes the mandatory services and it has overriding importance, because of assuring the operation of the railway network. The mandatory services comprise guarantee of traffic safety, giving assistance in the case of accident, dangerous and special goods transport.

On payment of the base charge the infrastructure manager provides the following services: access to the tracks, stations, public freight facilities, yards/depots, basic information to the public and assistance in case of accidents.

4.2. Charge of Additional Services

It includes services, which are needed for secure railway transport (traffic control, traffic security), the usage of the catenary, refuelling and the use of different services at the stations, yards, warehouses.

4.3. Extra Charges

Extra charges are justified in the following cases (over the base charge):

1. Additional trains on the heavy-traffic lines;
2. Use of Budapest's terminals for passenger trains;
3. Exceeding of the punctuality needed (set down in the services contracts);
4. Access to information and communication network;
5. Non-use of booked paths;
6. Depending on the time when the train is running (day, period of day).

4.4. Reductions

Infrastructure manager can give reductions in the following cases:

1. Taking advantage of the path for more years – on the basis of the contracts;
2. Use of the weakly frequented lines;
3. Running train at the same time (on all days of the week);
4. Shifting paths from saturated lines to alternative lines/services;
5. Shifting paths from peak to off-peak hours.

4.5. Repayments

If the infrastructure manager does not perform or incompletely complies with his obligation (that was undertaken in the 'services contract'; e.g. infrastructure failure, bad signalling or plan delays in all services) then railway operators can claim a penalty. If a railway operator's train breaks down and delays other services, then he pays penalty to the infrastructure manager and after that the railtrack pays penalty to other (innocent) operators. If the services run better than benchmark due to good network performance then the operators pay bonus to the infrastructure manager.

5. Conclusions

The fee is an index number, which expresses the quality of the railway infrastructure. Value of the charge shows how the service standard changes. The user charges for the railway infrastructure involve a direct feedback and they show how the increase/decrease of the state subsidy influences standard level of the rail transport [9].

The user charges have to be determined so that they demonstrate the condition of the national railway infrastructure – in comparison with EU-railways – and the national/foreign railway companies should be informed about the volume of the fee.

The following have to be done in the near future:

- Adjustment to EU Directives;
- The methodology should be implemented;
- Opening the Hungarian rail transport market.

References

[1] HODGSON, I., Railway Infrastructure Charges and the Commission Proposal 98 (480). *United Nations Economic Commission for Europe Trans European Railways (UN ECE TER) Seminar on User Charges for Railway Infrastructure*, July 7-9, 1999, Bratislava.

[2] User Charges for Railway Infrastructure, ECMT Round Table 107, Paris, 1998

[3] TÁNCZOS, K., The Condition System for the Efficient Operation of the Integrated Transport Infrastructure. Academic Dissertation, Budapest, 2000 (in Hungarian).

[4] TUCHHARDT, R., The New Infrastructure Charging System. Description of the System, *United Nations Economic Commission for Europe Trans European Railways (UN ECE TER) Seminar on User Charges for Railway Infrastructure*, July 7-9, 1999, Bratislava.

[5] FRAK, K., Infrastructure Access Charging Method, *United Nations Economic Commission for Europe Trans European Railways (UN ECE TER) Seminar on User Charges for Railway Infrastructure*, July 7-9, 1999, Bratislava.

[6] SALAVECZ, J., Information on Developments of User Charges for Railway Infrastructure in Hungary, *United Nations Economic Commission for Europe Trans European Railways (UN ECE TER) Seminar on User Charges for Railway Infrastructure*, July 7-9, 1999, Bratislava.

[7] FARKAS, GY., Modernisation of the 'Rules to Calculate Costs of Production in Railway Business', with Special Regard to Profitability Examination of the Running InterCity Trains between Budapest and Pécs. MSc. Thesis, June, 1996, pp. 63–68 (in Hungarian).

[8] BOKOR, Z., Practical Use of Controlling in Transport for the Case of Rail Transport. *Scientific Review of Transport* **10** (1999), pp. 368–376, and **12** (1999), pp. 451–458 (in Hungarian).

[9] FARKAS, GY., Financing Options of the Transport Infrastructure with Special Regard to Methodological Issues of the Defining of the User Charges for Railway Infrastructure, Thesis, February, 2000 (in Hungarian) pp. 76–89.

PERIODICA POLYTECHNICA SER. TRANSP. ENG. VOL. 28, NO. 1–2, PP. 117–122 (2000)

IDEAS REFLECTED IN THE FUTURE PUBLIC TRANSPORT ASSOCIATION IN THE AREA OF BUDAPEST

Katalin TÁNCZOS* and Barnabás KISS**

Department of Transport Economics
Faculty of Transportation Engineering
Budapest University of Technology and Economics
H–1521 Budapest, Hungary
Tel: 36 1 463 3265, 36 1 463 1061
Fax: 36 1 463 32 67
*e-mail: ktanczos@kgazd.bme.hu
**e-mail: bkiss@kgazd.bme.hu

Received: November 10, 2000; Revised: February 6, 2001

Abstract

Western countries with high motorisation realised years ago the fact that the excessive increase in the rate of motorisation would have to be restricted especially in densely populated urban areas. There have been enormous efforts to promote the public way of transport, which, however, can only be maintained with a highly professional structure and a management coupled with adequate financial provisions from the state. In the ex-socialist countries including Hungary, a great deal of change has been observed in the modal split towards the share of private cars. The increased use of private cars is threatening the dominance of the heritage of the past system, the rather low profile in terms of service standards but quite reliable in terms of timetable density and relatively cheap public transport. Following the examples of the West in this regard without paying attention to adjusting the specific characteristics in solutions of the problem can only bring about committing similar mistakes and thus paying an enormous price.

Keywords: Budapest Transport Association, urban public transport, setting fair tariffs.

1. Problems in Urban Transport in Hungarian Cities

It is only wise to learn from the mistakes of the West, i.e. of the almost total abandoning of public transport. On the other hand, where reasonable, cutbacks cannot be avoided. The clear solution seems to be a partly state-financed structure. We have had the opportunity for several years to closely look at the examples and to learn the lessons in France and in Germany [1] where public transport is totally market-oriented but partly state-financed.

Of course, the modal split in both countries is less favourable than in Hungary but due to enormous efforts it is on the increase and, moreover, the structure applied there deserves more attention. Our benefit of learning from the Western past could be a much more economic outcome of events and could also mean the retaining of the existing modal split values.

Beside the continuous decrease of modal split, a most important problem of the present Hungarian public transport sector, and in a way a reason for that, is the

very old rolling stock or more explicitly the lack of sufficient funds for the renewal. It must be examined what sort of state financing construction from what resources and the application of what kind of service standards will bring about the needed changes in the present situation.

The problems concern the country in various ways but the most interesting and also mostly affected areas are the urban centres and their catchment areas. Among them Budapest ranks the first, a capital and in many aspects a present and in many others a future regional centre in this Eastern part of Europe. The heritage of the past economic system is less and less able to cope with the dynamic challenges of increased quality standards and therefore, and partly due to the cancelling of certain lines and the decreasing of the timetable frequency of others, public transport is losing share. However, Budapest being a metropolis has the potential and the most chances to gradually switch to a market-oriented public transport service.

2. Justification of the Budapest Transport Association

As the Budapest Transport Company, the BKV is dominating the market it would be more favourable to change the quality standards of transport service from inside. The idea of the complex unified tariff system of the Budapest Transport Association, the BKSZ [2] could mean changes of the present situation. However, this integrated system will hopefully bring about positive future changes in the perceived utility of travellers, it will be implemented with many difficulties and through long debates, since all the parties [3] concerned will experience a decrease of tariff income at first.

It is of utmost importance to realise this fact and to elaborate methods for making up this 'lost income' and to identify possible resources. Elaboration of a more competitive tariff system than the present one is a must.

2.1. Specific Problems of Partners of the BKV

Debates on authority will result from the fact that the tariff policies of the three companies are dependent either on the municipality or the state. For example, the Volán companies still use cross financing for making up the inequalities between the profitability of their inter-city and city branches. While inter-city branches are free to set tariff levels and are thus profitable well explained by the rules of market economy, city lines struggle with losses resulting from their obligation to accept the decision of each city council to set tariffs at a socially acceptable level, which levels in many cases do not assure the conditions for profitability. The city council cannot decide on standards without granting funds. In this situation for the Volán companies the only choice is cross financing. And it is a certain share of the inter-city profit that will be the victim of the integration.

The situation is even worse if we look at what a small segment of the MÁV (Hungarian Railways) will be concerned by the changes induced by setting up the Budapest Transport Association. Therefore, any positive changes resulting from the new system will not coincide with the interests of the MÁV who will feel the adverse impacts rather than the positive changes.

In this regard, it is the BKV that should most enthusiastically emphasise the formation of the association, since the future operation territory of the BKSZ is the very territory of BKV's present activities and a good solution will contribute to the promotion of the BKV itself.

2.2. The Possible Scenario Regarding Tariff Changes

By applying in-depth investigation of existing tariff differences for each possible journey the company now setting the lower tariff will perceive that the elimination of the differences in price will adversely change the present excess of travellers and thus the turnover so far experienced. However, it must also be examined whether the balance of cost-benefit is positive or not for each line, since in the latter case more passengers will only increase the deficit.

So the reasoning based on the fact that a company attracting more travellers must in the new system be entitled to receive the major part of the income cannot be judged right or wrong in itself. Projecting present shares of low profile service to future rights of income from higher service standard activities is totally false. Instead, it should be examined which provider to what extent will add to the building up of higher service standards.

2.3. A Just Allocation of Income Based on Quality Output

As expected a pivot of debates will be the allocation of the income from the activities concerned. The allocation based on the share of use, i.e. on the result of simple traffic counts would seem evident for some parties but this simple technological-organisational method does not take into consideration the different cost-benefit ratios and the perceived utility of the passengers at each company at all (consult *Table 1* and *Fig. 1* where the thickness of the arrows shows the value of each input and output). However, this would be the most important point in the debate if we want to reject the still existing esprit and want to recur to the market-oriented way of thinking.

The most important fact is whether the travellers will experience better service standards in an aggregate way or not. In addition, no matter what the changes might be if there is no traffic volume increase, or at least the decrease does not slow down or even stop.

For a better understanding suppose that in the context of S (designating the complex technical, organisational and financial system of transport activities) there

are three providers X, Y and Z, using indices x, y and z, respectively or in general i?

- with allocated income I (Ix, Iy and Iz, respectively),
- with aggregated network performance p (px, py, and pz, respectively) in terms of passenger-kilometres,
- and with quality indicator q (qx, qy and qz, respectively), which is a function of factors such as the sum of perceived utility (U) of each passenger and the cost benefit ratio (CB) of the service, i.e. $q = f(U, CB)$.

Thus, there are two possible income allocation procedures/functions as described in *Table 1*, whilst the more just, second one is also detailed in *Fig. 1*:

Table 1. Allocation of income according to two different approaches

Allocation of income according to traffic counts	Complex allocation of income according to traffic counts and quality indicators
$Ii = f(pi)$	$Ii = f(pi) * qx/(qx + qy + qz)$

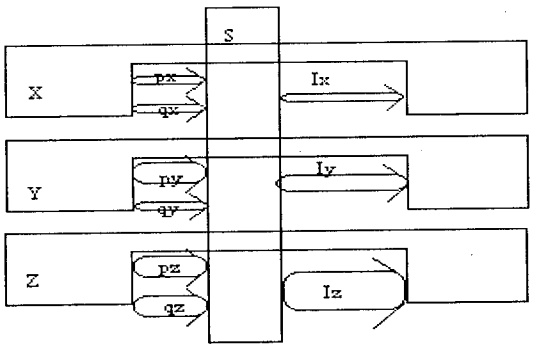


Fig. 1. Complex allocation of income according to traffic counts and quality indicators

The tariff levels will be set most probably near the present two or three different values. It is hard to imagine, however, that it will be at the present lowest level in spite of all synergic effects. The introduction of new prices with conditions rather unchanged cannot be made accepted by the users. A possible higher price level must be counterbalanced if not overwhelmed by the resulting synergic effect of joint systems and the user must perceive the improved quality in the service, such as more rational integrated timetables and reliable connections in reality.

3. State Funding with Strictly Activity-Specific Allocations

Is it wise to accept the above-mentioned different cost-benefit ratios for different modes of transport or is it more expedient advisory to try to modify these values so that they all will approximate a certain optimum level? It is essential for the solution that the activities be based on a market-oriented platform, which will implicitly mean that these figures will by themselves become optimal.

For the sake of avoiding abrupt increase in tariff levels due to the application of market rules it is expedient to apply the well-accepted rule in urban public transport of making profit with handling state funds as commercial income. Of course, the rate of the state's contribution must be controlled through strict and rational accounting.

Moreover, in order that the change of cost-benefit ratios does not concern the whole companies – since this way the positive processes launched in the local context would be hindered by the massive inertial potential of the companies in the case of MÁV and Volán – it is very much advised to allocate the amount needed directly to the Budapest-specific division of each company. State contribution already exists but without any transparency and the above mechanism would allow a responsible way of use of public funds in transport.

4. Conclusion

All these ideas are, however, subject to debate and can only be implemented fully if besides inter-company contracts changes will affect the legal context providing adequate conditions. The legal context has also to set certain directives for the companies and the authorities.

Remarks:

Note 1: The Department of Transport Economics has had extended relationships with the University of Karlsruhe (W. Rothengatter) and the Laboratoire d'Economie des Transports de Lyon (J. M. Cusset)

Note 2: It is expected that a unified tariff system based on an association incorporating the services of the *city transport* buses, tramways, trolleys, underground and light rails, the *national railways* and the *inter-city bus service* providers with their lines that run inside or in the agglomeration area would bring about positive synergic effects for both providers and travellers. The idea is mostly based on the presence of masses of people for whom commuting is an everyday reality. The actuality of the problem is backed by the phenomena of the urban sprawl which is more and more a reality for Budapest and the agglomeration.

Note 3: The three companies concerned are BKV as predominant service provider, the Hungarian Railways, MÁV, having radial connections to the city centre with

three railway head stations and the inter-city bus operator Volán company also running lines inside the heart of the city.

References

[1] A Budapesti Közlekedási Szövetség létrehozását megalapozó vizsgálatok – Zárójelentés (Final Report to the Examinations Aiming at the Implementation of the Budapest Transport Association) – 1996, TRANSMAN (Consulting for Transport System Management Ltd).

[2] Szempontok a fővárosi közlekedéspolitika elveinek és eszközrendszerének pontosításához (Aspects Aiming at a More Exact Methodology and Transport Policy of the Capital) – Dr. János MONIGL, 1998, TRANSMAN (Consulting for Transport System Management Ltd).

[3] Setting the Scene – The Case of Budapest, Central and Eastern European Forum – Prof. Dr. Katalin TÁNCZOS, 7-8 April 2000, Budapest.

[4] Gestion et Organisation Locales, Financement, Memento 1997 des Transport Publics Urbains – 1998, GART Groupement des Autorités Responsables de Transports.

[5] *The Urban Transport Crisis in Europe and North America* – John Pucher and Christian Lefevre, 1998, Macmillan.

PERIODICA POLYTECHNICA SER. TRANSP. ENG. VOL. 28, NO. 1–2, PP. 123–130 (2000)

THE USE OF MODERN SATELLITE SYSTEMS IN THE RAILWAY TRAFFIC

Gábor Kálmán KISS

Department of Transport Operation
Budapest University of Technology and Economics
H–1152 Budapest, Hungary
Tel.: (36 1) 432 38 99 Fax: (36 1) 342 93 95

Received: October 12, 2000

Abstract

In this paper the satellite-based railway positioning systems are summarised. In the first part the advantages of the satellite-based railway systems, short summary of the GPS system and the application today are given. Further on a system model is presented which can help us to survey the whole transport process.

Keywords: satellite, positioning, passenger and freight transportation, GPS-based train influence system.

1. Introduction

The rail passenger and freight transportation has lost in market position not only in Hungary, but all over Europe. The passengers and carriers increasingly prefer the road transportation. The growing mobilisation in the passenger transport service, the need for quick travel and not least the price level are the factors that influence the passengers in choosing between different means of transport. There is only one way for the national railway companies to make up their loss of market and backlog. They should undertake significant technological innovations.

The satellite-based telecommunication and positioning systems are in practical use both in the USA and in Japan. The next pages will deal with the advantages of satellite-based railway systems and their development options.

2. The Advantages of the Satellite-Based Railway Systems

- The maintenance expenses of the rail-track can be reduced leaving out the outdoor equipment installed along the tracks (e.g. the breaking of electric circuits).
- The 'compatibility errors' can be eliminated by connecting local systems.
- Using satellite-based systems, the traffic safety can be enhanced on secondary lines without interlocking equipment.

- The damages caused by vandalism can be reduced leaving out the outdoor equipment.
- The bottlenecks are recognisable so it is possible to solve them in time.
- Integrated combined transport systems can be created by means of which the transparency of the traffic flow of material and means of transport would be increased.
- The introduction of 'Moving Block' (dynamic track sections) results in economic and environmental advantages. The traction parameters are improved (less traction energy consumption, reduced noise and dust pollution thanks to the better planned braking process).
- Large areas (even more countries) can be covered.
- Quick build up.
- Exact positioning and control of moving vehicles.

3. The GPS System

The NAVSTAR-GPS (**NAV**igation **S**ystem with **T**iming **A**nd **R**anging-**G**lobal **P**ositioning **S**ystem) has been developed for military use by the US Defence Ministry. The system – including 26 active satellites – ensures optimal covering in all corners of the Earth. The accuracy of the 'time and distance measuring' based system has increased to 4 meter since 1st January 2000 when Bill Clinton President of the US discontinued the SA (Selective Availability).

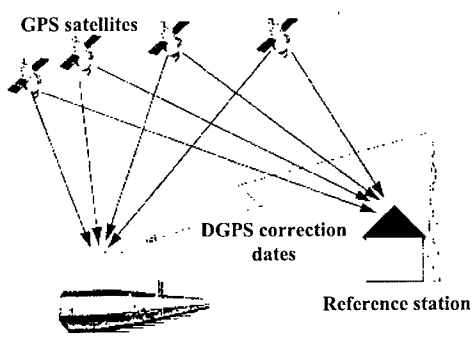


Fig. 1. DGPS system components

In the 90s there was in use a procedure that helped to reach the cm precision. It was called Differential GPS (DGPS). The fundamental idea was to transmit correction data from a reference station (*Fig. 1*), of which position was well known. The GPS receiver could work (depending on the duration and the speed) with correction data in the interval of 10–100 centimetres.

ALCATEL

Secondary Line Signalling System

Level 5:
- Points with electric drive.
- Inductive signals (balises) can be used in case of intensive traffic.
- AULI: It sets the points and other track-objects in the right direction automatically.
- RZÜ: Course-diagram drawing method.
- Automatic train-influence equipment and speed-signalling display

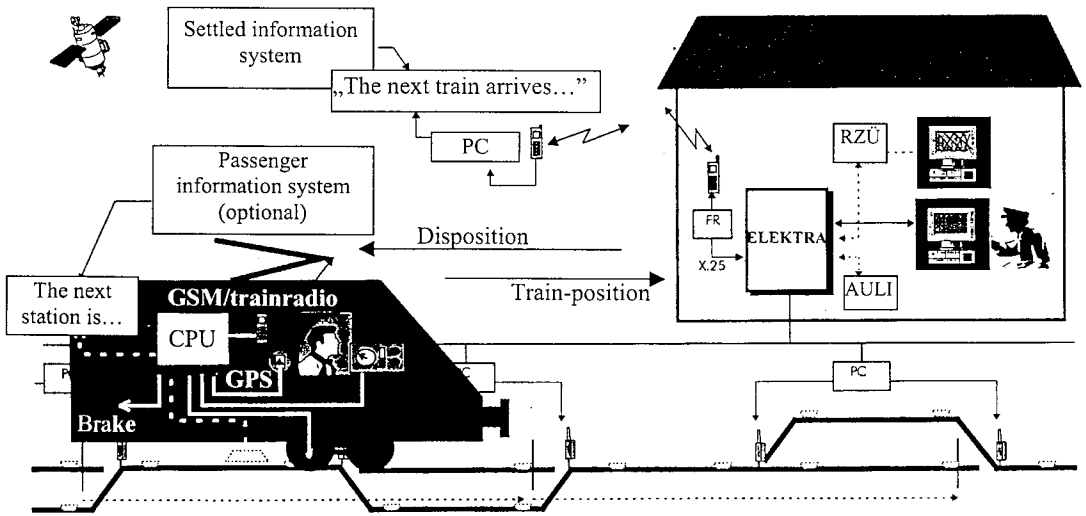


Fig. 2. ALCATEL's Secondary Lines Signalling System

4. Satellite-Based Railway Systems

The number of satellite-based railway systems is considerably behind the number of road transport systems. Their enumeration is easier and more obvious if we use the following 4-group collecting.

In the *first group* there are the GPS-based geodesy procedures. They are the first members of the satellite-based railway systems. They were firstly applied in territories where high accuracy was needed; however, the existence of security justification of the system was not among the requirements. Thanks to the GPS-based geodesy equipment the accuracy of route allocation can be improved and the survey of the existing routes can be done faster. This pioneer equipment is called SURVER [1] developed by the engineers of the DB AG. In 1994 this special field of geodesy got to the point that the measurements could be done with higher accuracy in a few hours. Earlier for the same work it was needed even more days.

In the *second group* there are the GPS-based train influence systems. They make possible to leave out the traditional track-electric-circuits. The following Austrian and Japanese applications are just a few examples in addition to the American systems (e.g. the Union Pacific Railway Companies Positive Train Separation [2]).

The Austrian ALCATEL Company has developed a new technology to control the transport on secondary lines carrying little traffic [3]. The GPS-based system guarantees a higher level of safety than the 'inter-station' system and its build up expense is also favourable.

The Secondary Lines Signalling System (SLSS) helps the work of traffic controllers because there is no need to give and ask for permission and the system also documents itself automatically.

To reduce the build up expenses engineers use spring points to allow crossing of trains. It is possible to save the operating and maintenance costs of cable laying equipment by the abandonment of the signals and track-electric-circuits. On industrial tracks ALCATEL uses key-locked points which can be opened and closed, if necessary, on the spot by the engine driver.

The SLSS is designed to work on 5 different levels (*Fig. 2*). The easiest function is the location of the train position and its display on the monitor installed in the Control Centre. The other utmost point is the GPS-based Train Command Centre. In this application the traffic controller is able to give dispositions to the trains or even the timetable-based computer can command the whole line. ALCATEL uses a train radio to exchange data between the control centre's computer and on-board equipment. The current speed or brake commands can bee seen on a monitor placed in the driver's cab so the traditional signals – along the track – can be left out.

The Japanese railway also has a GPS-based security system, called CARAT (Computer And Radio Aided Train Control System) [7].The main function of this system is to protect the maintenance group on the tracks. Each leader of these groups has a warning equipment. This mobile receiver gets the 'own position' data from the coming train. So the workers can leave the track safely before the train's

arrival. The advantages of the system are the higher security and the lower expenses because the engine driver does not have to reduce the speed and increase it again. This safety equipment is on trial operation on the Joetshu-Shinkansen line.

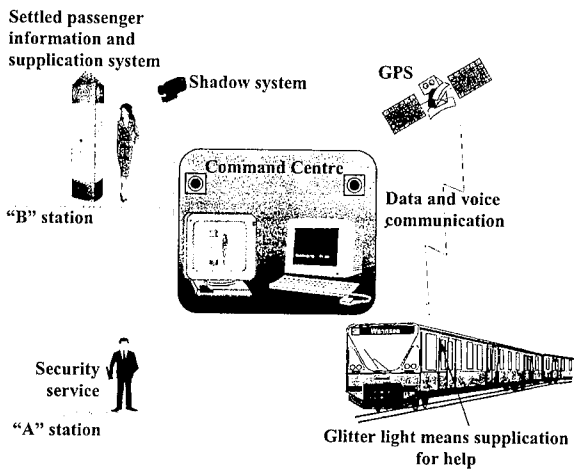


Fig. 3. S-Bahn Berlin-System overview

In the *third group* there are the railway transport systems. The most important and interesting application is the German mobile equipment, the **Sa**telliten-Ha**ndy** (Sandy) [4]. The basic idea was to prevent the damages of the goods that in Hungary reach up to 8%. These damages are caused not only by stealing and pilfering, but mostly by the shunting crew who are working with old technology and sometimes irregularly.

The SANDY can be mounted on the carriage (of course it is thief proof). Through its sensors the equipment is able to collect the parameters continuously and transfer them to the system operator via SMS (Short Massage System) or by train radio every 5–10 minutes. It is possible to adjust in advance the threshold values of the sensors. When the measured value is higher than the threshold value, SANDY will send an 'ALARM' message to the operator. After this message the necessary steps can be done.

Besides the satellite-based transport systems there is another application, the modern GPS-based passenger information systems constituting the *fourth group*. In Berlin (Germany) the S-Bahn supplements this information system (*Fig. 3*) with a communication system [5].

The passengers use it only in emergency situations. They either go to the settled passenger information and supplication box or ask for help on the board of the S-Bahn's train by pressing a button. The security service is regulated from the Command Centre to the GPS determined place.

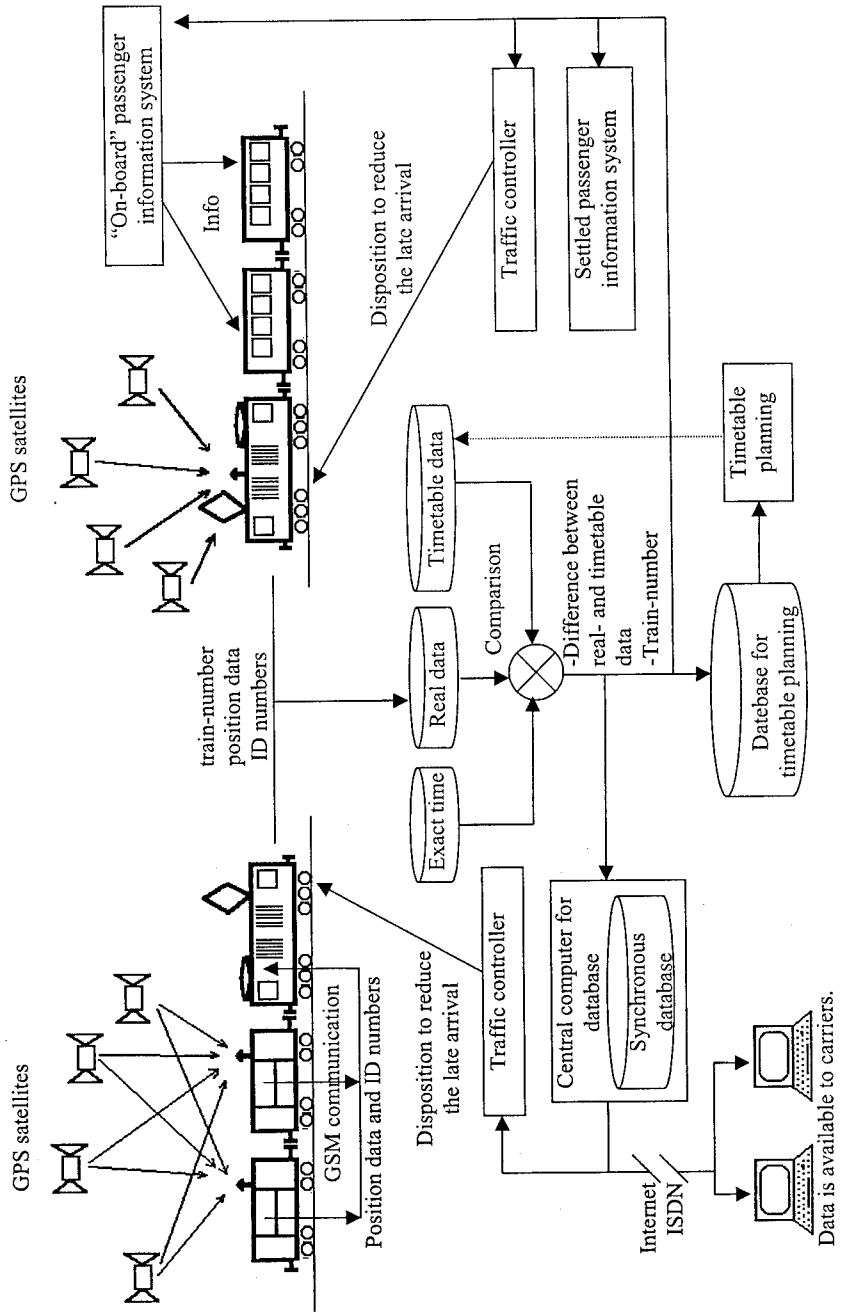


Fig. 4. Informatogram of the GPS-based passenger and freight transportation

The system works normally as an information system, naturally based on GPS positioning data.

5. Development Options

Today the rail passenger and freight transport systems are working isolated from each other. It is necessary to create a system model [6], which can help us to survey the whole transport process. This is the only way to see all the opportunities that satellite-based railway systems offer.

For *passenger transportation*, the satellite-based positioning offers the most in the field of passenger information.

The GPS-based *settled passenger information system* can automatically show the precise late arrival instead of the information nowadays '10 minutes late', or '15 minutes late'. This is the only way to increase the feeling of security, the calculableness and hereby the trust in the railway.

The *on-board information system* has a 3-step development option. In the first step a satellite-based Central Unit is connected to the waggons' tape recorders and loudspeakers. The name of the next station, the remaining time to the next stop and the possible late arrival can be automatically told to the passengers. The service level can be improved through the change of the information communication (e.g. using TV screen, or LCD monitor) or making it frequently. Possessing a database it is possible to inform the passengers about the name of the stations and sights along the track by using interactive monitors (*Fig. 4*).

Another important application of the satellite-based passenger systems is timetable planning. All information systems need the time differences between the timetable and the real data. By storage of these time differences we can create a database for timetable planning. This makes the work of the planning engineer easier and more precise.

The railway *freight transportation* has two different groups.

The first one is the '*one group train*', where all carriages have the same destination. The GPS followed object is only the engine, the carriages are followed just logically. Their positions can be established with a precision of 600–1000 meters (because of the greatest train length). Considering that the exact positioning data is not so important for the carriers it is allowed to follow just the train and not every carriage. The only weakness of the system is the human factor. For connecting the carriages logically to the GPS-based engine it is necessary to know their identification numbers. Before the departure the numbers have to be collected and transmitted (by radio) to the system. After that you can ask for the position of the carriages through the engine's position. When the train stops and when a carriage is uncoupled several times and later sent on with another train the risk of data missing is increasing. Besides the monitoring method should be strictly followed.

The second group of the rail freight transportation is the '*carriages following*

concept'. It means that every waggon has its own GPS-based identification equipment, that can avoid data missing during the 'radio guided carriages data collecting'. The waggons with their ID numbers can be operated much easier.

The mobile equipment sends SMS messages to the engine, so it is cost-effective. The data from the engine will be transmitted to the traffic controller by using train radio contact. The time difference, i.e. the deviation from the timetable is representable on maps. Using the 'synchronous database' the traffic controller can be informed on the late arrival and so he is able to make a decision in time. On the basis of this information the rail freight transportation could be more automatic, precise and more accurate (e.g. statistics, valuations).

The data of transport are available for carriers. They can follow their freight in 'real time'. (The earlier systems followed the carriages just like cross-section. There was no possibility for continuous positioning registration. The carriers known only that the waggon was 'under transport', or it was 'being unloaded').

6. Summary

This paper offers a survey of the wide area of satellite-based railway positioning systems. The purpose was to outline an overall image and to systematize the applications. The keywords are the GPS-based geodesy procedure, the GPS-based train influence system and the satellite-based railway transport and information system. The different areas need different levels of safety information and positioning data, but all the applications are designed to put the railway forward, to stabilize its position on the transport market.

The satellite systems have a lot of advantages and their costs are immediately recovered. We hope that the satellite-based railway technologies have a great future.

References

[1] LUDWIG, S., GPS in the Railway Geodesy. (In German). *Eisenbahningenieur*, **9** (1997).
[2] SAUER, S. J., Burlington Northern Santa Fe GPS Survey Project. (In English). American Railway, 1997.
[3] HARTBERGER, M., Train Influence System by ALCATEL. (In German). Eisenbahnsicherungstechnik, 1999.
[4] The Little SANDY is Very Big in Collecting of Data. (In German). *Deine Bahn*, **8** (1999).
[5] JACOB – RENKEN, Passenger Security and Satellite-Based Positioning in S-Bahn Berlin. (In German). *SIGNAL+DRAHT*, **3** (1996).
[6] WESTSIK, GY., *Transportation Information Technology and Telematics*. (In Hungarian.) Műgyetemi Kiadó, 1997.
[7] GRÜLLER, R., NAVSTAR-GPS, the Global Positioning System. (In German.) *Eisenbahn-Revue*, **5** (1997), **6** 1997, **1–2** (1998).
[8] KLATT, K.: The Possibilities of the Use of GPS in the Passenger Information and Railway Line Control. (In German.) SIGNAL+DRAHT, **3** (1998).

PERIODICA POLYTECHNICA SER. TRANSP. ENG. VOL. 28, NO. 1–2, PP. 131–142 (2000)

LQG/LTR CONTROLLER DESIGN FOR AN AIRCRAFT MODEL

Balázs KULCSÁR

Department of Control and Transport Automation
Budapest University of Technology and Economics
H–1111, Budapest, Bertalan L. u. 2., Hungary
Fax: +36-1-463-3087
e-mail: kulcsar-b@kaut.kka.bme.hu
Phone: +36-1-463-3089

Received: Nov. 10, 2000

Abstract

The automation of the operation of aircraft equipment lightens and diminishes the manipulation of the pilot during the flight. This article gives some introduction about the fundamental aspects of the LQG/LTR control theory, and it shows an application option through an example. The movement of an aircraft can be modeled with a linear time invariant dynamic system, which must be controlled by a flight controller. This article contains the synthesis and analysis of the stability and other qualitative control parameters of the flight control.

Keywords: LQG/LTR controller, aircaft.

1. Introduction

In this paper we shall give a short view of the so-called Linear Quadratic Gaussian theory, which can be consulted for more details in [1] and [2]. KWAKERNAAK and SILVAN, ANDERSON and MOORE, DAVIS and VINTER, ASTRÖM and WITTENMARK, FRANKLIN and POWEL and many others worked on this theory. Then we revise the main stages of a Linear Quadratic Gaussian/Loop Transfer Recovery method, which was elaborated by Doyle and Stein [16]. This article dicusses an example of an aircraft flight controller design, using first LQR and LQG, afterwards the LQG/LTR methods. We can refer to [3], [4] and [5], [7] where we can find some basic applications for Linear Quadratic controller design for simplified aircraft models.

First of all we will examine the traditional optimal controller design of the aircraft flight controller system using LQR (Linear Quadratic Regulator) and LQG (Linear Quadratic Gaussian) method. With LQG/LTR method we recover the stability margin of the Kalman filter at the plant output.

In the LQG case we can use the separation principle, which means that we are able to design the LQG controller in two steps. First, the design of the LQR (Linear Quadratic Regulator), and then we have to find a state estimator, an LQE (Linear Quadratic Estimator) applying a modified cost function.
where x is the state vector, r is the referential signal, y is the output vector, $w(t)$ is the state noise, $v(t)$ is the sensor noise, $\hat{x}$ is the estimated state vector, e is the error

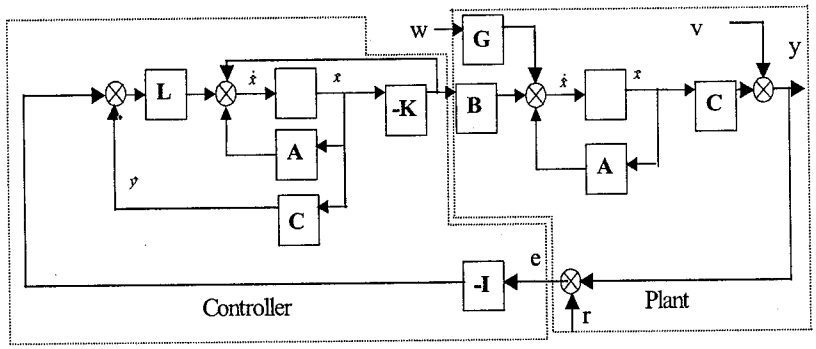


Fig. 1. The LQG controlled plant

signal, $\hat{y}$ is the estimated output vector, A is the state matrix, B is the input matrix, C is the output matrix, G is the disturbance input matrix, K is the static feedback gain matrix, L is the stationary Kalman filter static gain matrix, the observer gain, I is the identity matrix.

There exist, in our case, two types of noises (external, internal). We will calculate with an external stochastic noise (for example the air turbulence), and an internal random measurement noise. The two random (external disturbance and sensor) noises are white, Gaussian (normal) zero-mean stationary vector processes. We know the covariance matrices, and the external disturbance noise has a disturbance input matrix G. To find a controller which stabilizes our plant, we have to solve a stochastic integral problem.

2. Problem Setup

As we know well, both the LQ regulator, and the Kalman filter have good robustness and performance, so the LQG controller would have good properties too. Unfortunately this is not the case.

We have two methods for the LQG design. Either we design the feedback gain before the design of the Kalman filter, or we make first a filter and a feedback gain after. We receive a different solution if we use the two different methods of the LQG design.

Are we able to find a way for tuning the LQG controller with a parameter? The answer is yes. There is a way of recovering either the full state-feedback properties or the state estimator properties.

3. Loop Transfer Recovery (LTR) at the Plant Output

The state space representation of the plant is given by

$$\begin{aligned} \dot{x} &= Ax + Bu + Gw, \\ y &= Cx + v, \end{aligned} \tag{1}$$

and its transfer function is the following

$$G(s) = C(sI - A)^{-1}B, \tag{2}$$

where $G(s)$ is the Laplace transform of the transfer function from control input to the output. The state equations of the LQG controller become

$$\begin{aligned} \dot{\hat{x}} &= A\hat{x} + Bu - L(r + \hat{y}), \\ \hat{y} &= C\hat{x}, \\ u_{\text{opt}} &= -K\hat{x}, \end{aligned} \tag{3}$$

and the controller transfer function is

$$G_C(s) = K(sI - A + BK + LC)^{-1}L. \tag{4}$$

The controlled LQG closed loop can be written as

$$\begin{aligned} \varepsilon &= x - \hat{x}, \\ \dot{x} &= Ax + Bu = Ax - BKx + BK\varepsilon, \\ \dot{\varepsilon} &= Ax + Bu - A\hat{x} - Bu + L(r - C\hat{x}) = (A - LC)\varepsilon + Lr, \\ \begin{bmatrix} \dot{x} \\ \dot{\varepsilon} \end{bmatrix} &= \begin{bmatrix} A - BK & BK \\ 0 & A - LC \end{bmatrix} \begin{bmatrix} x \\ \varepsilon \end{bmatrix} + \begin{bmatrix} 0 \\ L \end{bmatrix} r, \\ y &= [C\ 0] \begin{bmatrix} x \\ \varepsilon \end{bmatrix}, \end{aligned} \tag{5}$$

where: ε is the state error or the difference between the real and the estimated state. The transfer functions of the open loop for input is

$$G_H(s) = G_C(s)G(s) \tag{6}$$

and for the output

$$G_H(s) = G(s)G_C(s). \tag{7}$$

We shall design and recover the return ratio of Kalman filter. The stability margins of the Kalman filter are guaranteed to be good. Now, the key of the recover is the Q weighting matrix.

First, we will write the open loop transfer function

$$G_H(s) = (C(sI - A)^{-1}BK(sI - A + BK + LC)^{-1}L. \tag{8}$$

Using the notations

$$\begin{aligned} \Phi(s) &= (sI - A)^{-1}, \\ \Psi(s) &= (sI - A + LC)^{-1} \end{aligned} \tag{9}$$

and applying the matrix inversion lemma

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(DA^{-1}B + C^{-1})^{-1}DA^{-1} \tag{10}$$

one obtains for the open loop transfer function

$$\begin{aligned} G_H(s) &= C\Phi(s)BK(\Psi(s)^{-1} + BK)^{-1}LG_H(s) \\ &= C\Phi(s)BK[\Psi(s) - \Psi(s)B(K\Psi(s)B + I))^{-1}K\Psi(s)]L \\ &= C\Phi(s)BK\Psi(s)[I - (BK\Psi(s) + I)^{-1}BK\Psi(s)]L \\ &= C\Phi(s)BK\Psi(s)(BK\Psi(s) + I)^{-1}L. \end{aligned} \tag{11}$$

The feedback gain K for LQ design is given by

$$K = R^{-1}B^T P, \tag{12}$$

where $P \geq 0$ is the maximal solution of the Control Algebraic Ricatti Equation (CARE)

$$AP + PA^T + Q - PBR^{-1}B^T P = 0. \tag{13}$$

Suppose that the weighting matrix Q can be written as

$$Q = Q_0 + qM. \tag{14}$$

We will examine the situation while $q \to \infty$:

$$\lim_{q\to\infty}\left(\frac{AP}{q} + \frac{PA^T}{q} + \frac{Q_0}{q} + M - \frac{PBR^{-1}B^T P}{q}\right) = 0 \tag{15}$$

and [16]

$$\lim_{q\to\infty}\frac{P}{q} = 0. \tag{16}$$

If $q \to \infty$ K can be obtained from the equation

$$\begin{aligned} M &= \frac{PBR^{-1}B^T P}{q}, \\ q^{1/2}M^{1/2} &= R^{-1/2}K, \\ \lim_{q\to\infty}(K) &= \lim_{q\to\infty}(q^{1/2}R^{-1/2}M^{1/2}). \end{aligned} \tag{17}$$

Choosing $M = I$! the open loop transfer function becomes

$$\begin{aligned}\lim_{q\to\infty}(G_H(s)) &= \lim_{q\to\infty}(C\Phi(s)Bq^{1/2}R^{-1/2}\Psi(s)(Bq^{1/2}R^{-1/2}\Psi(s)+I)^{-1}L) = \\ &= \lim_{q\to\infty}(C\Phi(s)BR^{-1/2}\Psi(s)\Psi^{-1}(s)R^{1/2}B^{-1}L), \qquad (18)\\ \lim_{q\to\infty}(G_H(s)) &= C\Psi(s)L.\end{aligned}$$

This result says that with a q very high we can approach the return ratio of the Kalman filter.

The first table represents the effect of the LQG loop transfer recovery.

Table 1. Loop transfer recovery

The poles of the $G_C(s)_{q\to\infty}$	$\to$	The zeros of the plant $C\Phi(s)L$
The zeros of the $G_C(s)_{q\to\infty}$	$\to$	The zeros of the $C\Phi(s)B$

The stages of the design LQG/LTR for output

1. We must find the required Kalman filter gain L.
2. Afterwards we determine the LQ feedback gain, using the substitutions: $M = I$, $Q = Q_0 + qM$. We want to find the K feedback gain while $q \to \infty$.

Remarks

1. We must take care of the type of LQG/LTR tuning at the plant input, because we enhance the state noises while we approximate the required LQ gain. Our aim with increasing q is to find a compensated plant which converges sufficiently closely to (18) in a large range of frequency.
2. In the case of a non-minimal phase system the optimal loop cannot be rebuilt, because the zeros of the open loop transfer function are not all negatives.

4. Linear Quadratic Gaussian/ Loop Transfer Recovery Design at the Plant Output of a Hypothetical Fighter Aircraft

During the LQG/LTR controller design it is supposed that our aircraft is cruising at a constant altitude, and with a constant velocity. The linearized longitudinal equations are simple, ordinary linear differential equations with constant coefficients. The coefficients in the differential equations are made up of aerodynamic stability derivatives, mass and inertia characteristics of the aircraft.

The longitudinal dynamic model of the fighter aircraft, in linearized form equations is [17]:

$$
\begin{aligned}
\dot{q}(t) &= -0.8q(t) - 0.0006u(t) - 13.2\alpha(t) - 19\delta_E(t) - 2.5\delta_f(t),\\
\dot{u}(t) &= -0.014u(t) - 16.64\alpha(t) - 32.2\theta(t) - 0.66\delta_E(t) - 0.5\delta_f(t),\\
\dot{\alpha}(t) &= q(t) - 0.0001u(t) - 1.65\alpha(t) - 0.16\delta_E(t) - 0.6\delta_f(t),\\
\dot{\theta}(t) &= q(t),
\end{aligned}
\tag{19}
$$

where for control inputs $\delta_f(t)$ is the perturbed flapperon angle deflection, $\delta_E(t)$ is the perturbed elevator angle deflection, and for states $\alpha(t)$ is the perturbed angle of attack, $\theta(t)$ is the perturbed pitch angle, $q(t)$ is the pitch rate, $u(t)$ is the perturbed horizontal velocity.

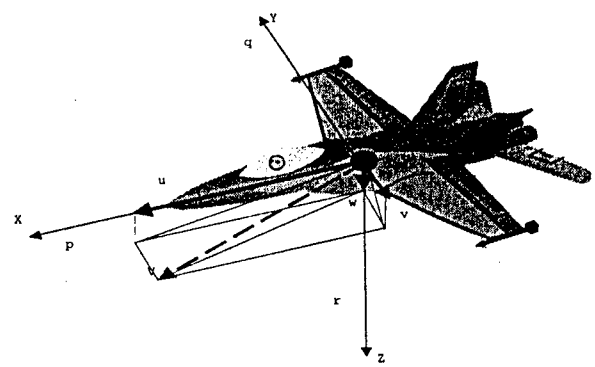


Fig. 2. The aircraft body coordinate system (ABC).

We can write these equations in state-space form:

$$
\begin{bmatrix} \dot{q} \\ \dot{u} \\ \dot{\alpha} \\ \dot{\theta} \end{bmatrix} = \begin{bmatrix} -0.8 & -0.0006 & -13.2 & 0 \\ 0 & -0.014 & -16.64 & -32.2 \\ 1 & -0.0001 & -1.65 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} q \\ u \\ \alpha \\ \theta \end{bmatrix} + \begin{bmatrix} -19 & -2.5 \\ -0.66 & -0.5 \\ -0.16 & -0.6 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \delta_E \\ \delta_f \end{bmatrix},
\tag{20}
$$

$$
\begin{bmatrix} \theta \\ \gamma \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 1 \end{bmatrix} + \begin{bmatrix} q \\ u \\ \alpha \\ \theta \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \delta_E \\ \delta_f \end{bmatrix},
$$

where γ is the flight path angle $(\theta - \alpha)$.

4.1. Time Domain Analysis of the Uncontrolled Aircraft

Before we design compensator for our hypothetical aircraft dynamics it is expected to analyse the pure dynamics. In this section we study the transient response to specific test signals. The step function and Dirac's delta or impulse function are chosen as testing inputs. We can conclude some information about the transient behaviour. We neglect disturbance. We want to control the pitch angle with the elevator deflection (input 1 – output 1), and the flight path angle will be controlled with the variation of the flap (input 2 – output 2).

The impulse and step response functions can be seen in *Fig. 3*. The short and long period (phugoid) oscillations of the uncontrolled aircraft dynamics are two special characteristics of the aircraft movement. They are caused by the special placement of the poles. In this case we have $p_1 = -1.1512 + 3.4464i$, $p_2 = -1.1512 - 3.4464i$, $p_3 = -0.0058 + 0.0264i$, $p_4 = -0.0058 - 0.0264i$. The poles, which are close to the imaginary axes, simulate the long period movement, and the other pole-pair conjugated causes the short period dynamics. The periods are readily obtained once the eigenvalues are known. The duration of the short period movement is $t_s = 1.826$ s and the phugoid is $t_L = 241.6$ s.

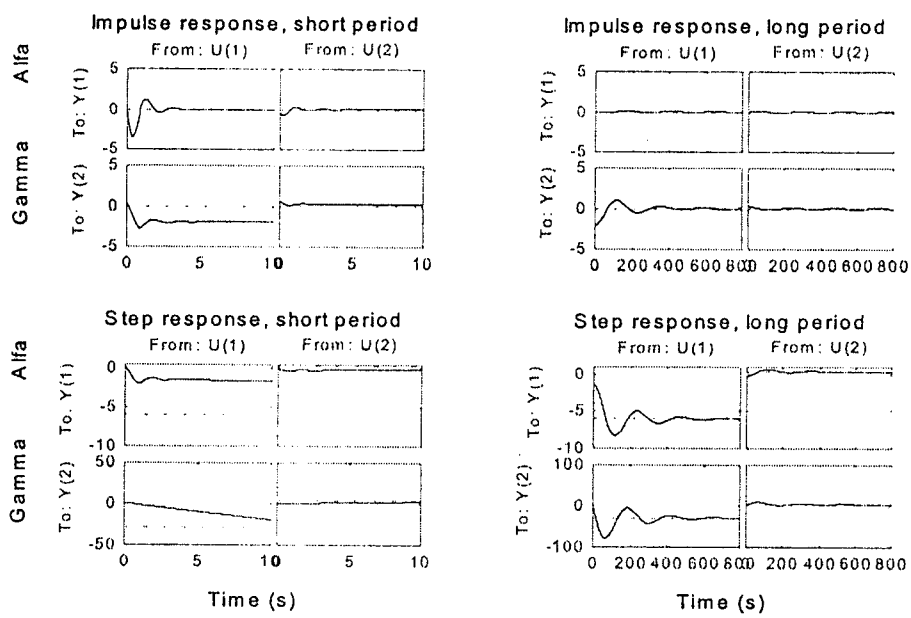


Fig. 3. The behaviour of uncontrolled dynamics in time domain

4.2. LQR Controller Design for a Hypothetical Aircraft

Our linear system is controllable and observable, and because of the equality of the Kalman rank of the observability and controllability matrix it is minimal too. We are going to design a LQR controller and we suppose that all states can be measured for the first approach. We are able to compare the results of the pure LQ controller with a LQG controller (see the next section), and see how the stochastic observer modifies the LQ results.

$$\begin{aligned} &\operatorname{rang}(O_4(C, A)) = \operatorname{rang}[C\ CA\ CA^2\ CA^3]^T \\ &= \operatorname{rang}(C_4(A, B)) = \operatorname{rang}[B\ AB\ A^2B\ A^3B] = \dim x(t) = 4. \end{aligned} \tag{21}$$

We will use the Inverse Square Rule [4] for determining the weighting matrices Q and R. Also we can write and see the solution in *Fig. 4*:

$$Q = \begin{bmatrix} 0.0365 & 0 & 0 & 0 \\ 0 & 0.000025 & 0 & 0 \\ 0 & 0 & 8.2 & 0 \\ 0 & 0 & 0 & 0.672 \end{bmatrix}, \qquad R = \begin{bmatrix} 3.6 & 0 \\ 0 & 3.6 \end{bmatrix}. \tag{22}$$

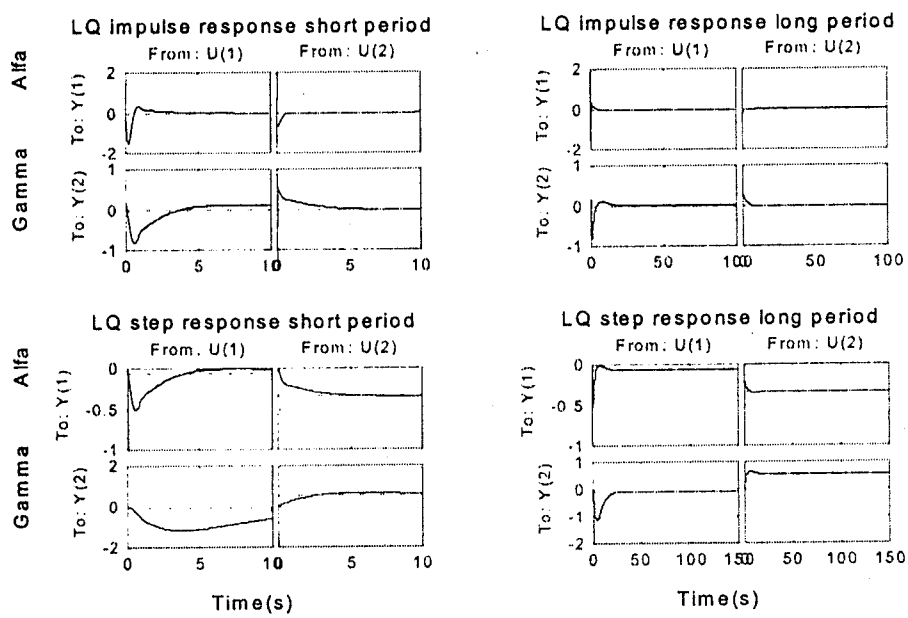


Fig. 4. The behaviour of LQ controlled dynamics in time domain

4.3. LQG Controller Design for a Hypothetical Aircraft

We cannot analyse the dynamical behaviour of our aircraft without external and measurement noise. We suppose an external stochastic turbulent air and an internal noise at the same time.

We will expand the Linear Quadratic Regulator problem to the Linear Quadratic Gaussian problem including the noises. The Separation Principe helps us to design in two different stages the LQG controller. First of all, we shall utilise the weighting matrices Q and R. As for the covariance matrices V and W of the Kalman filter I propose:

$$W = \begin{bmatrix} 0.01 & 0 & 0 & 0 \\ 0 & 0.01 & 0 & 0 \\ 0 & 0 & 0.01 & 0 \\ 0 & 0 & 0 & 0.01 \end{bmatrix}, \qquad V = \begin{bmatrix} 0.01 & 0 \\ 0 & 0.01 \end{bmatrix}. \tag{23}$$

When we compare the two solutions (LQR in *Fig. 4* and LQG in *Fig. 5*) received by the two different types of controller, we are able to conclude that the Kalman filter always modifies the final solution (*Fig. 4*).

The time simulation of the error function (difference between the real and the estimated state) gives us a convenient result. The error function returns to zero in about 4 s from the initial value 1.

Let us examine the behaviour of the model in frequency domain. *Fig. 6* shows how the open-loop singular values of the uncontrolled, LQG controlled, and LQR controlled loop plant change. In each case the two peaks of the maximal singular value, appropriate to the two couples of complex-conjugated pole are clearly visible.

4.4. LQG/LTR Controller Design for a Hypothetical Aircraft

The aim of the design is to recover the return ratio of the Kalman filter at the plant output. When we design the loop transfer recovery at the plant output method we want to get back the initial diagram of singular values of the open loop transfer function of the observed plant in a relatively large frequency band.

We have already calculated the Kalman filter gain L for the LQG controller. Also the first stage of the design is ready. In the second step increasing the factor q we approximate the initial observed plant. If we choose $q = 1$ the recovery is not acceptable only in a strictly narrow interval of frequency till about 0.07 rad/s. The $q = 3$ factor allows the recovery of the Kalman filtered plant in a rather wide interval (until 0.5 rad/s). Finally, $q = 10$ fits the limit 1 rad/s.

Perfect recovery should not be obtained in the case of a non-minimal-phased system. It is true that the LTR procedure can be applied to unstable plants without difficulty, since unstable poles are shifted by the feedback into the left half-plane

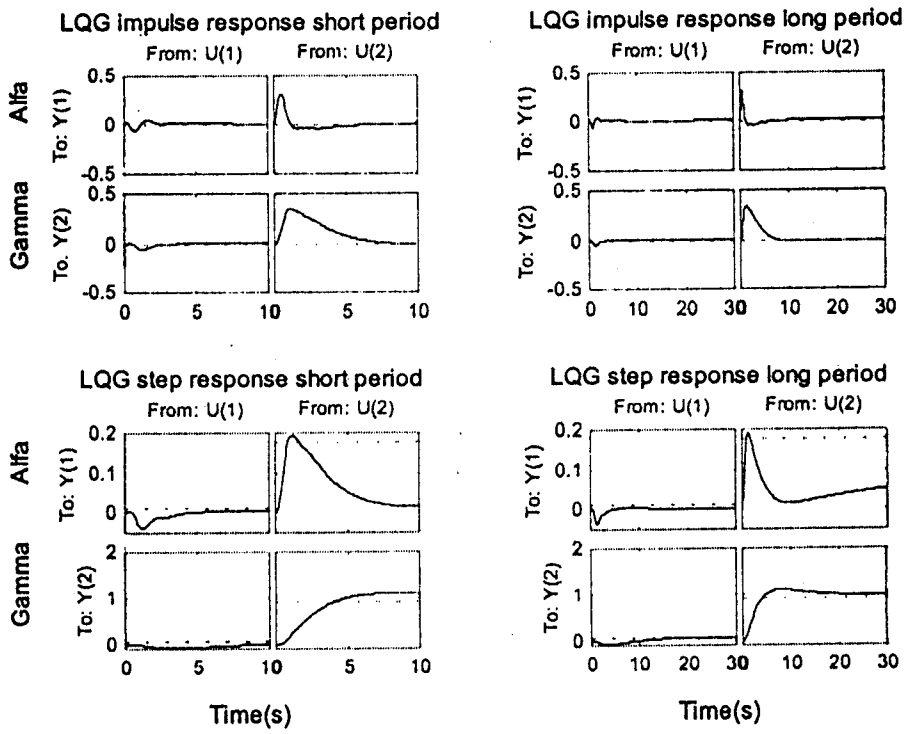


Fig. 5. The behaviour of LQG controlled dynamics in time domain

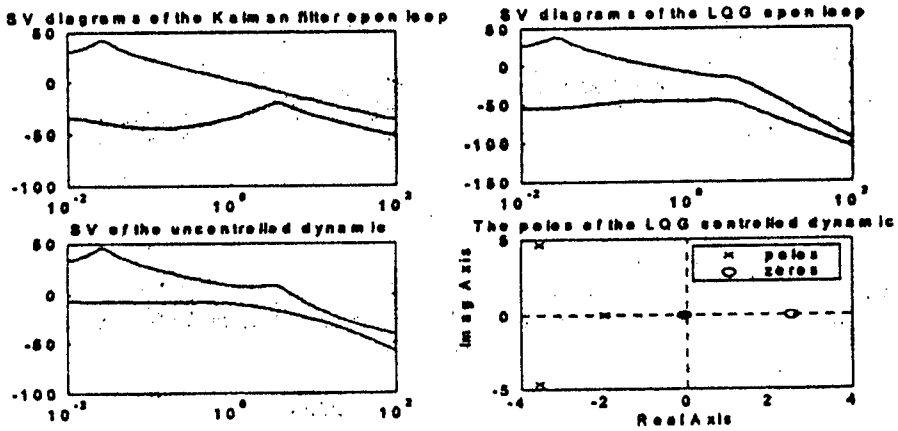


Fig. 6. The behaviour of uncontrolled and controlled dynamics in frequency domain

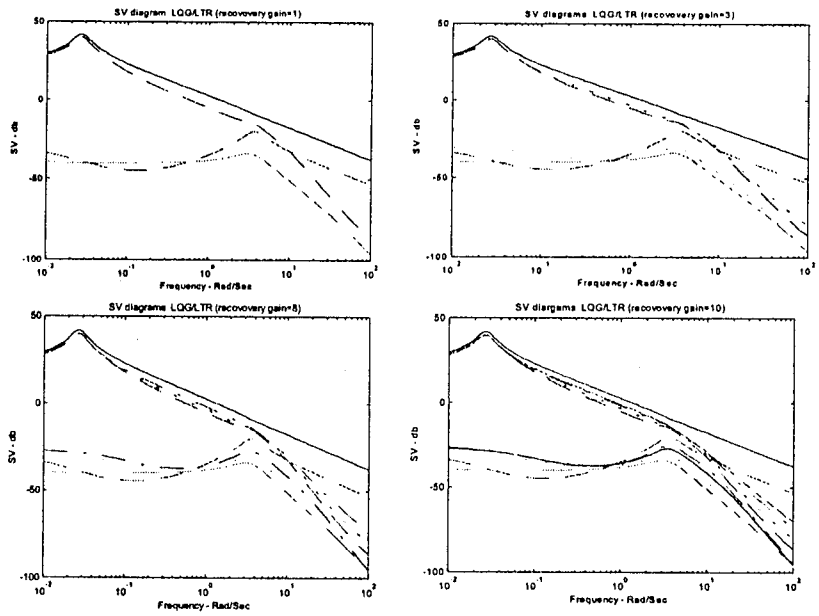


Fig. 7. The behaviour of LQG/LTR controlled dynamics in frequency domain

if all zeros and poles are in the negative real part. In our case the system is non-minimal-phased. So the recovery may be achievable at those frequencies at which the plant's response is very close to that of a minimal-phase plant. We conclude to accept the final stage of recovery when $q = 10$ because it assures the bandwidth of the referenced system. The behaviour of the LQG/LTR in time domain can be seen in *Fig. 8.* We can see the step response function, which is given by the controlled dynamic. Normally one wants to see the return of the step response function to the unit or to zero (unit from input 1 to output 1, and from input 2 to output 2, and in the other case return to zero). We have the output error when we use LQ controller. Also we have to design a referenced signal following controller called LQ servo.

5. Summary

In this paper we analyzed the aircraft dynamic described by a linear equation of the motion in time and frequency domain. We could see the effect of uncontrolled dynamics, and the design LQR, LQG and LQG/LTR controller. With LTR having obtained, we recovered a satisfactory return ratio for the Kalman filter, stochastic state-observer, we succeeded in recovering the return ratio at the plant output. The author utilized the MATLAB and CONTROL Toolbox to obtain these results of the simulations.

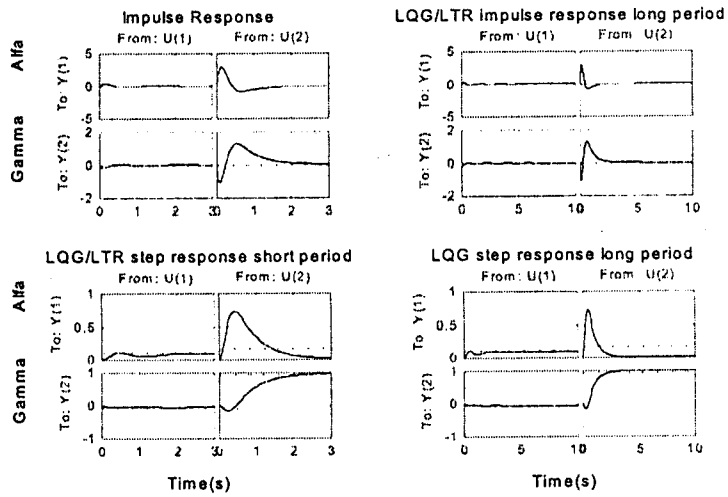


Fig. 8. The behaviour of LQG/LTR controlled dynamics in time domain

References

[1] MACIEJOWSKI, J. M., *Multivariable Feedback Design*. Addison-Wesley Publishers Ltd., 1989.
[2] BROGAN, W. L., *Modern Control Theory*, Pretince Hall, 1991.
[3] MCLEAN, D., *Automatic Flight Control Systems*. Prentice Hall, London, 1990.
[4] NELSON, R. C., *Flight Stability and Automatic Control*. Prentice Hall, London. 1998.
[5] ETKIN, B. – REID, L. D., *Dynamics of Flight – Stability and Control*. John Wiley & Sons, Inc, Toronto, 1996.
[6] ZHOU, K. – DOYLE, J. C. – GLOVER, K., *Robust and Optimal Control*, Prentice Hall, London.
[7] SZABOLCSI, R., Robust Controller Synthesis for the Aircraft Pitch Altitude Control System. *Repüléstudományi Közlemények, Szolnok*, Vol. XII. 29. No. 2000/1 pp. 79–88. (In Hungarian).
[8] KULCSÁR, B., Synthesis and Analysis of the Aircraft Flight Control Systems. *Repüléstudományi Közlemények*, Szolnok, Vol. XII. 29. No. 2000/1 pp. 91–101. (In Hungarian).
[9] STEIN, G. – ATHANS, M., The LQG/LTR Procedure for Multivariable Feedback Control Design. *IEEE Transaction on Automatic Control*, AC-32, (1987), pp. 105–114.
[10] DOYLE, J. C., Guaranteed Margins for LQG Regulators. *IEEE Transaction on Automatic Control*, AC-23, (1978), pp. 756-757.
[11] DOYLE, J. C. – STEIN, G., Robustness with Observers. *IEEE Transaction on Automatic Control*, AC-24, (1979), pp. 607–611.
[12] DOYLE, J. C. – STEIN, G., Multivariable Feedback Design: Concepts for a Classical/Modern Synthesis. *IEEE Transaction on Automatic Control*, AC-26, (1981), pp. 4–16.
[13] SZABOLCSI, R. – GÁSPÁR, P., Flight Control System Synthesis for the Aeroelastic Jet Fighter Aircraft, *Proc. of the 2nd IFAC Symp. on Robust Control Design, ROCOND 97*, Budapest, pp. 453–458, 1997.
[14] GÁSPÁR, P. – MICHELBERGER, P. – BOKOR, J. – PALKOVICS, L. – NÁNDORI, E., Iterative LQG Controller Design in Uncertain Parameter System, *Proc. of the 2nd IFAC Symp. on Robust Control Design, ROCOND 97*, Budapest, pp. 483–488, 1997.
[15] GÁSPÁR, P. – SZÁSZI, I., Robust Servo Control Design using Identified Models, *Proc. of the 3rd IFAC Symposium on Robust Control Design*, Prague, Czech Republic, 2000.
[16] KWAKERNAAK, H. – SILVAN, R., *Linear Optimal Control Systems*. New York, Wiley, 1972.
[17] ATHANS, M., Lecture notes. MCS-II, 1990.

PERIODICA POLYTECHNICA SER. TRANSP. ENG. VOL. 28, NO. 1–2, PP. 143–152 (2000)

EXPERIMENTAL RESEARCH ON LASER–MATERIAL INTERACTION

Xiufeng WANG*, János TAKÁCS**, György KRALLICS***, András SZILÁGYI** and Tamás MARKOVITS**

*Beijing University of Aeronautics and Astronautics
100083 Beijing, China
** Department of Manufacturing and Repair of Vehicles
***Department of Mechanical Technology and Materials Science
Budapest University of Technology and Economics
H–1521 Budapest, Hungary

Received: Oct. 1, 2000

Abstract

This paper presents a new method of real–time measuring technology to investigate the thermo–physical process of laser–material interaction during laser non–melting surface processing. In the thickness direction of the specimen, the temperature distribution was measured with Thermovision Infrared System in real–time. The dynamic micro–deformation of the specimen was studied using laser beam reflex amplifier system. Experiment results display the thermo–physical process of laser–material interaction and lay a foundation of further researching on the process of laser processing.

Keywords: laser–material interaction, temperature distribution and deformation distribution

1. Introduction

Laser processing technology is an advanced and highly efficient manufacturing method. It has been applied in airplane industry, defense industry, automobile industry, mechanical industry and material industry, etc. Due to extremely high laser output power density (more than 10^4 W/cm^2), extremely quick heating speed and self–cooling speed ($10^3 - 10^6$ °C/s) on the material surface, the process of laser–material interaction is regarded as a very complex thermo–physical process under the interaction between temperature, phase transformation and stress–strain. It relates to laser, material, thermal–physics, dynamics and many other scientific fields. After the interaction of laser beam with material, material deformation is the result of thermal deformation and transformation deformation cooperation. Final stress of material interior is superposition of thermal stresses and transformation stresses. When stresses there exceed the elastic limit of material, material will produce plastic deformation. If material is thinner such as sheet metal, it will get beneficial plastic deformation, which is laser sheet metal bending. If material is thick enough, it will not cause obvious plastic deformation, this stress is kept in the material and regarded as residual stress. Residual stress will have a great influence on the usage properties of the material. We expect that the residual compressive

stress is produced on the material surface in order to improve bearing capacity of material.

Laser transformation hardening on the surface material and laser sheet metal bending are two different application fields in laser non–melting surface processing. Although experts have done lots of research on theory and experiment [1]–[16], due to lacking research on the thermo–physical process of laser–material interaction, the behaviour and regularity of strain–stress during the process have not been understood clearly. The technological parameters chosen by trial and error cannot ensure the optimal quality of treated surface and get the shape and direction of deformation expected.

This paper presents a new method of real–time measurement technology to investigate the thermo–physical process of laser–material interaction during laser non–melting surface processing. In the thickness direction of the specimen, the temperature distribution was measured with thermovision infrared system in real–time. The dynamic micro–deformation of the specimen was studied using laser beam reflex amplifier system. Experiment results reveal the thermo–physical process of laser-material interaction and lay a foundation of further researching on the process of laser processing.

2. Experimental Method

The experimental principle scheme for laser–material interaction is shown in *Fig. 1*.

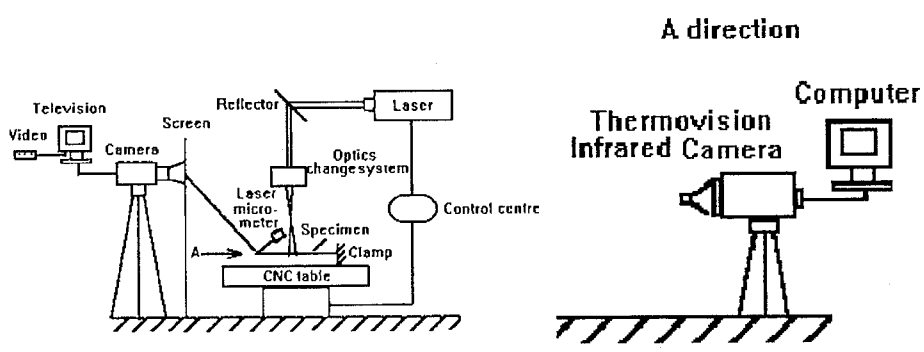


Fig. 1. The experimental principle scheme for laser–material interaction

A 1800 W CO_2 laser is used in continuous wave mode, its power density distribution is TEM_{00}. The wavelength of laser beam is 10.6 μm, focus length of lens is 254 mm. The laser beam is guided through an optical fiber cable from the controller to the laser beam head. On/off switch, the output power and pulse duration of it, and the movement of the $x - y$ table are controlled by CNC–program. CNC control unit accuracy is 0.005 mm. The specimen is clamped at one end as a cantilever beam and put 21 mm under focal point of lens (beam diameter on the

specimen surface is 2 mm). The specimen is irradiated in the middle by the laser beam, the continued pulse duration is 0.782 s. A small piece of the plexiglas is stuck to its upper side of free end. Laser micrometer irradiates on the plexiglas and reflects to the screen stuck by a piece of coordinate paper in order to amplify specimen deformation. A camera used at a frequency of 24 Hz records the process of movement of light point during laser–material interaction. The television connected with the video displays the movement of light point recorded by camera. At the same time the thermovision infrared camera used at an image frequency of 6.25 Hz records the real–time temperature distribution in the thickness direction of the specimen in the heated region. Computer displays images recorded by the thermovision infrared camera.

The typical die material C45 (100 mm × 2 mm × 2 mm) and low carbon steel sheet St14 (100 mm × 2 mm × 2 mm, 100 mm × 2 mm × 1.5 mm) are chosen. Their chemical compositions are shown in *Table 1*.

Table 1. Chemical compositions, [%]

	C	Si	Mn	P	S
St 14	0.08	0.03 – 0.1	< 0.4	< 0.025	< 0.025
C45	0.42 – 0.50	0.15 – 0.30	0.40 – 0.70	< 0.04	< 0.04

The surface of material is sprayed with a graphite coating (Graphite 33) in order to enhance absorption of energy.

3. Experiment Results

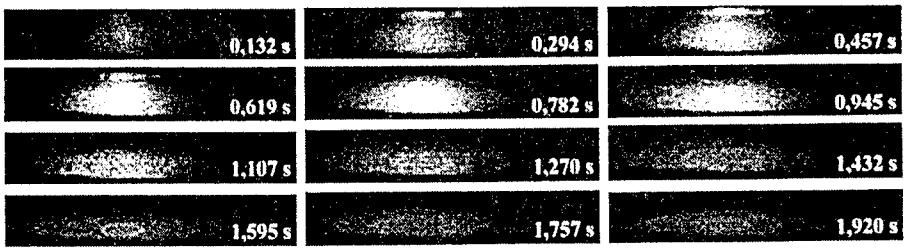
3.1. The Temperature Distribution

1. During the process of laser–material interaction, the temperature distribution images in the thickness direction of the specimen in the heated region are shown in *Fig. 2* and *Fig. 3*.

Fig. 2 and *Fig. 3* indicate the variation of temperature distribution during the process of laser–material interaction. This process comprises two phases, the heating phase and the cooling phase. When laser irradiates the specimen surface, light energy changes into heat energy in a very short time and then goes into the specimen very quickly. When laser beam switches off, the specimen will cool very quickly. But due to heat transfer, the effect of boundary can be seen obviously: heat energy transferring to the lower side surface is reflected into the specimen.

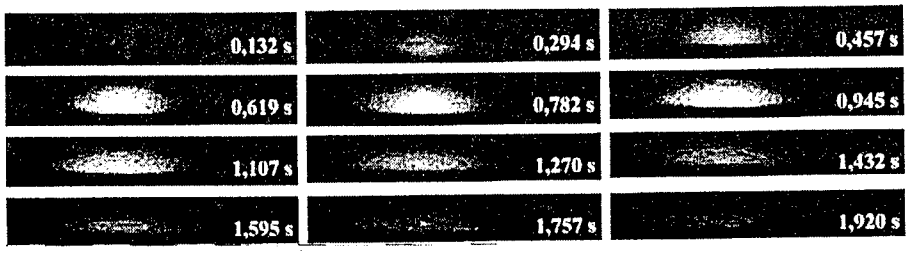
At the end of pulse duration, the temperature distribution images in the thickness direction of the specimen in the heated region are shown in *Fig. 4* and *Fig. 5*.

Fig. 4 and *Fig. 5* indicate the distribution of maximum temperatures recorded during the process of laser–material interaction. It is obvious that because the



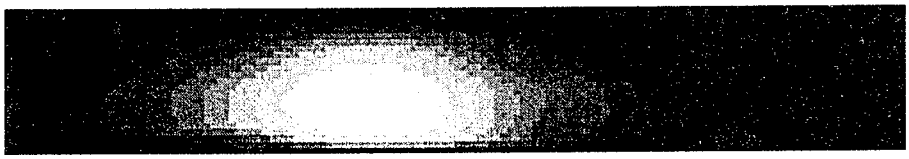
Material: St14; Thickness: 2 mm; Laser output power: 100 W; Time of process: 1.92 s

Fig. 2. Variation of the temperature distribution during laser–material interaction



Material: C45; Thickness: 2 mm; Laser output power: 100 W; Time of process: 1.92 s

Fig. 3. The variation of the temperature distribution during laser–material interaction



Material: St14; Thickness: 2 mm; Laser output power: 100 W; Pulse duration: 0.782 s

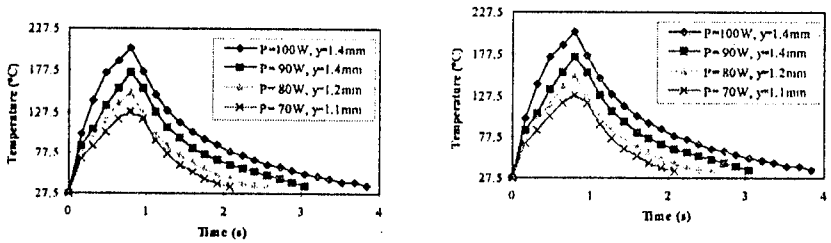
Fig. 4. The temperature distribution image at the end of pulse duration

Material: C45; Thickness: 2 mm; Laser output power: 100 W; Pulse duration: 0.782 s

Fig. 5. The temperature distribution image at the end of pulse duration

specimen thickness is finite, heat energy conducted to the lower side surface is reflected into the specimen and makes temperature of specimen thickness direction in one position reach the maximum. This position of the maximum temperature depends on the energy absorbed by specimen.

2. During the process of laser–material interaction, the variation of temperature versus time in the specimen thickness direction is shown in *Fig. 6*.



Material: St14; Thickness: 2 mm
Pulse duration: 0.782 s

Material: C45; Thickness: 2 mm
Pulse duration: 0.782 s

Fig. 6. The variation of temperature versus time

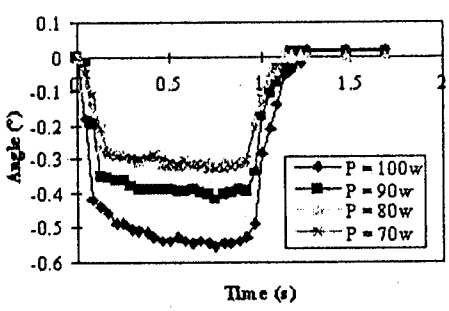
Fig. 6 indicates that the recorded maximum temperature increases with the laser output power improvement, but the position of the maximum temperature is different. For material St14, the position of the maximum temperature moves toward the lower side. This means that the more energy is absorbed by material, the larger the depth of the heat conductance inside the material is. At the same time, the more the heat energy that conducts to the lower side surface, is reflected into the specimen, the nearer to the lower side the position of the maximum temperature recorded is. However, for material C45, the position of the maximum temperature is almost constant. This is the reason why the phase transformation occurs in the material. In the pulse duration, the austenitic transformation needs absorb the heat. The more the energy absorbed by material, the more the heat absorbed by the austenitic transformation is.

3.2. The Dynamic Micro–Deformation Distribution

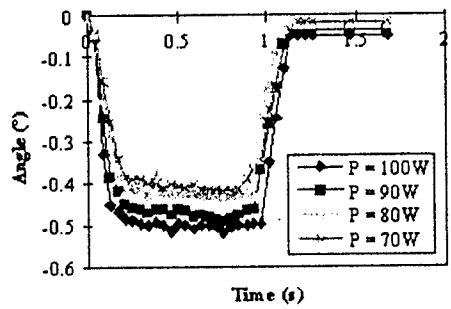
The variation of the dynamic micro–deformation versus time during laser–material interaction is shown in *Fig. 7*.

Fig. 7 indicates when the laser beam irradiates the specimen surface, the specimen deforms away from the laser beam because of thermal expansion.

With continual pulse duration, the deformation is larger and larger. But due to heat conductance, heat energy that conducts to the lower surface is reflected into the specimen and makes temperature of the specimen interior improve. This decreases the deformation and makes the curve show a constant value for a short time after the pulse duration finishes. After laser beam switches off, because the temperature



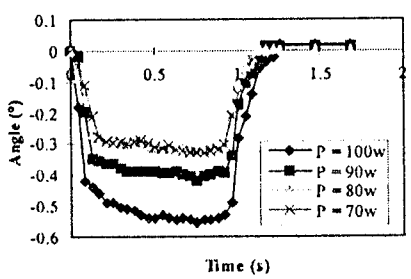
Material: St14; Thickness: 2 mm
Pulse duration: 0.782 s



Material: C45; Thickness: 2 mm
Pulse duration: 0.782 s

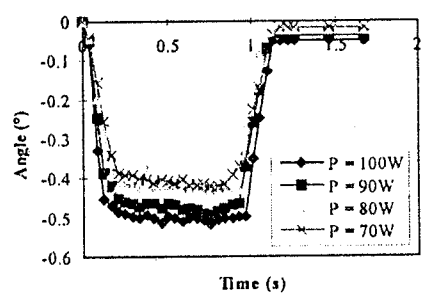
Fig. 7. The variation of the dynamic micro–deformation versus time during laser–material interaction

is falling, the specimen contracts and the deformation of the specimen away from laser beam reduces gradually. For material St14, the specimen deforms towards the laser beam at last. But for material C45, due to the martensitic transformation, volume expansion decreases the deformation towards the laser beam and makes the specimen deform away from the laser beam finally. The value and direction of final deformation depend on the energy absorbed by the specimen. It is obvious that the higher the laser output power, the larger the deformation.



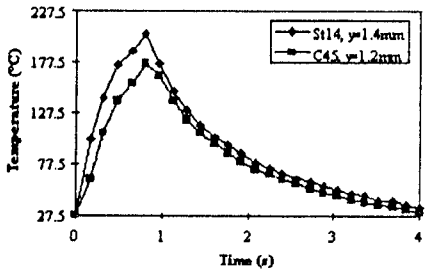
Material: St14; Thickness: 1.5 mm
Pulse duration: 0.782 s

Fig. 8. The influence of laser output power on the temperature distribution



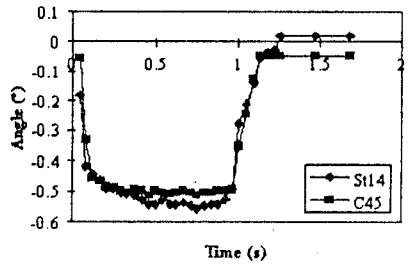
Material: St14; Thickness: 1.5 mm
Pulse duration: 0.782 s

Fig. 9. The influence of laser output power on the dynamic micro–deformation



Thickness: 2 mm
Laser output power: 100 W
Pulse duration: 0.782 s

Fig. 10. The influence of the composition of material on temperature distribution



Thickness: 2 mm
Laser output power: 100 W
Pulse duration: 0.782 s

Fig. 11. The influence of the composition of material on dynamic micro–deformation

4. Discussion and Conclusion

4.1. The Influence of Technological Parameters on Laser–Material Interaction

The main factors influencing laser–material interaction are laser output energy, the composition peculiarities of the material and the specimen thickness. Their effects are expressed as different temperature distributions and different dynamic micro–deformation distributions. Experiments display the variation of temperature and dynamic micro–deformation versus time during laser–material interaction under the condition of different laser output power, material and specimen thickness. The detailed analyses are shown as follows.

1. The influence of laser output power on the temperature distribution and the dynamic micro–deformation are shown in *Fig. 8* and *Fig. 9*.

Fig. 8 indicates that the higher the laser output power is, the higher the maximum temperature recorded is. That means that the energy absorbed by the specimen determines the temperature value of the specimen. In addition, the position of the maximum temperature moves toward the lower side due to improvement of the depth of the heat transferring into the material interior and the heat energy that conducts to the lower side surface is reflected into the specimen.

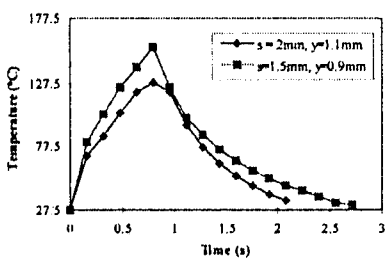
Fig. 9 indicates that the higher the laser output power is, the larger the specimen deformation during laser–material interaction is. This is the reason that the value of thermal expansion depends on the temperature distribution in the specimen. However, the specimen deformation depends on the value of thermal expansion.

2. The influence of the composition of the material on the temperature distribution and the dynamic micro–deformation are depicted in *Fig. 10* and *Fig. 11*.

Fig. 10 indicates that under the same technological parameters, the position of the maximum temperature of material C45 is nearer to the upper side. This is the reason that the phase transformation occurs in the material. In the pulse duration, the austenitic transformation needs absorbing the heat.

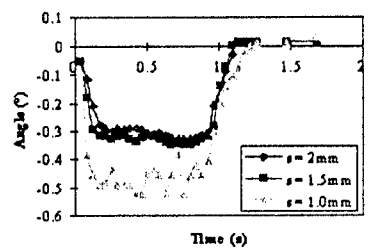
Fig. 11 reveals that under the same technological parameters, the composition of the material has a great influence on the final deformation of the specimen. For material St14, the specimen deforms towards the laser beam at last. But for material C45, due to the martensitic transformation, volume expansion decreases the deformation towards the laser beam and may cause the specimen deformation away from the laser beam, finally.

3. The influence of the specimen thickness on the temperature distribution and the dynamic micro–deformation is demonstrated in *Fig. 12* and *Fig. 13*.



Material: St14;
Laser output power: 70 W
Pulse duration: 0.782 s

Fig. 12. The influence of the specimen thickness on temperature distribution



Material: St14;
Laser output power: 70 W
Pulse duration: 0.782 s

Fig. 13. The influence of the specimen thickness on dynamic micro–deformation

Fig. 12 shows that under the same technological parameters, the thinner the specimen is, the more the heat energy is that transfers into the lower surface and is reflected into the specimen. So the corresponding temperature is higher. At the same time, the relative position to the thickness of the maximum temperature is nearer to the lower side.

Fig. 13 indicates that under the same technological parameters, the thinner the specimen is, the larger the specimen deformation is. This is the reason that the thinner the specimen is, the less the section moment of the specimen is. So the specimen produces the more deformation under the same energy interaction. But if the specimen is thin enough a strong vibration is caused during laser–material interaction. Experiment result will be affected.

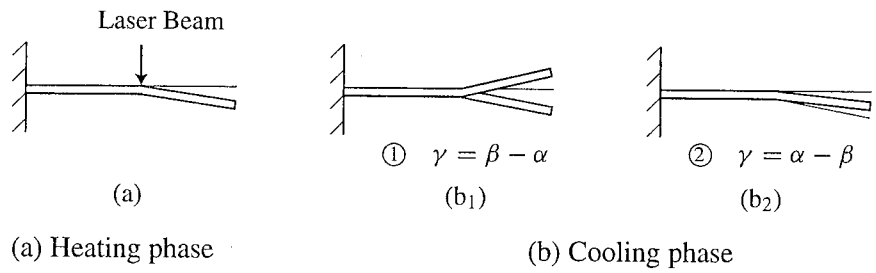


Fig. 14. The thermo–physical process of laser–material interaction

4.2. The Thermo–Physical Process of Laser–Material Interaction

Experimental result shows that the thermo–physical process of laser–material interaction is divided into heating phase and cooling phase.

- Heating phase
  When the laser beam irradiates on the specimen surface, the temperature on the upper surface in the heated region increases rapidly (not exceeding melting point of material). Strong temperature gradients mainly occur in the thickness direction of the specimen. Due to large thermal expansion and low yield stress on the upper surface at a high temperature, the material heated region produces compressive plastic deformation and causes material to accumulate. So the specimen deforms away from the laser beam. With a continual pulse duration, the deformation is larger and larger. But due to heat conductance, heat energy that conducts to the lower surface is reflected into the specimen and makes the temperature of the specimen interior improve. This decreases the deformation and makes the curve of specimen deformation show a constant value for a short time after the pulse duration finished. The angle of deformation is α as shown in *Fig. 14*(a). It depends on the temperature gradient in the thickness direction and the geometry constraint of the specimen.
- Cooling phase
  After laser beam switches off, the temperature on the upper surface is falling very quickly. Material contracts so that part of accumulative material recovers and compressive stress decreases. With the development of cooling process, due to heat conductance, the lower layer of the specimen expands continually and makes it elongate. So the deformation away from laser beam reduces gradually. As far as the material possesses phase transformation property, the martensitic transformation can cause volume expansion so that the deformation away from the laser beam increases. Therefore, the angle β of deformation is the result of thermal deformation and transformation deformation cooperation.

Thus, it can be seen that the final deformation angle of the specimen can be expressed by $\gamma = \beta - \alpha$. When $\beta > \alpha$, the specimen deforms towards the laser

beam as shown in *Fig. 14*(b_1). When $\beta < \alpha$, the specimen deforms away from the laser beam as shown in *Fig. 14*(b_2).

Acknowledgements

The authors gratefully acknowledge support for this study from the Department of Mechanical Engineering Technology and the Department of Materials Science and Engineering. They also thank Dr. Antal Sklánitz and Dr. Zoltán Dudás for their assistance during the experiments.

References

[1] SINDO KOU, D. – SUN, K. – LE, Y.P., The Fundamental Study of Laser Transformation Hardening. *Metallurgical Transactions A*. **14A** (1983), April, pp. 643–653.
[2] LISA, P. et al, The Importance of Thermal Stresses and Strains Induced in Laser Processing with Focused Gaussian Beams. *J.Appl. Phys*. **64** (1988), (11), 1 December, pp. 6274–6286.
[3] VAN BRUSSEL, B. A. – DE HOSSON, J. TH. M., Residual Stresses in the Surface Layer of Laser-Treated Steels. *Materials Science and Engineering*. **A161** (1993), pp. 83–89.
[4] BERGMANN, H. W. – DAMASCHEK, R., Laser Surface Heat Treatment of Tools. *Proceedings of the LANE'94*, eds. M. Geiger, F. Vollertsen, Meisenback Bamberg (1994), pp. 280–294.
[5] ZOON, K.K. – KIM, W. B, – NA, S. J., Shape Deformation of a Piston Ring Groove by Laser Surface Hardening. *Surface and Coating Technology* **78** (1996), pp. 157–167.
[6] WANG, H. G. et al, A Study of Thermal Stresses during Laser Quenching. *Journal of Materials Processing Technology* **63** (1997), pp. 550–553.
[7] FERNANDEZ, M. R., Steel AISI 1045: *Superficial Hardening by Laser and Tribological Behavior. Surface Modification Technologies*. The Institute of Material, London (1997) pp. 617–622.
[8] STEEN, W.M., *Laser Material Processing*. Second Edition, Springer, 1998.
[9] TAKACS, J., The Laser Technology – Technology for the Future. *Proceedings of First Conference on Mechanical Engineering*. Technical University of Budapest (1998), pp. 515–519.
[10] GEIGER, M. – VOLLERTSEN, F. – DEINZER, G., Flexible Straightening of Car Body Shell by Laser Forming. SAE paper no. 930279 (1993), pp. 354–361.
[11] GEIGER, M., Synergy of Laser Material Processing and Metal Forming. *Annals of the CIRP* **43/2** (1994), pp. 563–570.
[12] VOLLERTSEN, F., Mechanisms and Models for Laser Forming. *Proceedings of the LANE'94*, eds. M. Geiger, F. Vollertsen, Meisenback Bamberg (1994) pp. 345–359.
[13] VOLLERTSEN, F., Potentials of Laser Forming. *Laser and Electron*, (1996) pp. 34–37.
[14] HENNIGE, T., Laser Forming of Spatially Curved Parts. *Proceedings of the LANE'97*, eds. M.Geiger, F.Vollertsen, Meisenback Bamberg (1997), pp. 409–420.
[15] DOVC, M. – MOZINA, J. – KOSEL, F., Pulsed Laser Bending of a Plate as an Optodynamic Process. *Proceedings of the LANE'97*, eds. M. Geiger, F. Vollertsen, Meisenback Bamberg (1997), pp. 421–430.
[16] XIUFENG, W. – XIUYAN, W. – LI, W., Synergy of Laser Bending. *Proceedings of the 8th National Conference on Forging* (1999) pp. 100–106 (in Chinese).

INFORMATION FOR AUTHORS

Scope. Periodica Polytechnica Transportation Engineering is an international journal supported by the Budapest University of Technology and Economics. It accepts research and application oriented papers in the area of transportation engineering. The journal has a free electronic version (`www.pp.bme.hu/te`).

Initial submission. Manuscripts should be submitted in English in duplicate to the office of Periodica Polytechnica (see inner front cover) and should also retain a copy everything submitted. Authors are encouraged to send to `perpol@eik.bme.hu` an additional printable electronic copy (PostScript, PDF,...) for a possible quicker electronic review. Manuscripts are accepted for *review* on the understanding that the same work has not been and will not be either published or submitted elsewhere and that all the authors have given their approval for the submission of the paper.

Electronic submission. Periodica Polytechnica uses the typesetting system LaTeX. Therefore, authors are asked to submit their contribution *after acceptance of the manuscript* in LaTeX either in an email attachment or on floppy disk. If the authors use other typesetting or word processing system, we advise them to save and send the material in different formats (e.g. RTF, TXT, DOC,...) for easier convertibility.

Manuscripts. A correctly compiled manuscript should contain the following items: *title*; if it is long, an *abbreviated title* for running head on the odd pages, consisting no more than 50 characters (including spaces) and having no abbreviations; name(s) of *author(s)*; *affiliation and addresses* (institution, firm, mail address, e-mail, URL); an informative *abstract* of not more than 250 words; 3–6 *keywords*. Since the abstract and keywords of an accepted paper will be disseminated electronically, it should make minimal use of mathematical symbols and should not contain displayed formulas or any reference to the article. The abstract and keywords will be offered for databases and internet search engines.

A *separate sheet* must be attached to the manuscript with the title and the names of authors and with the name, mail address, e-mail address, office phone and fax numbers of the contact author.

Rules for LaTeX papers. Please use the standard article style with the standard packages (`array`, `amsmath`, `graphicx`, ...). Avoid visual editing to improve the appearance of the manuscript like adding extra space, forced page or line break. Please keep formatting macros to a minimum, to make conversions of files more accurate. Please use standard LaTeX commands for sectioning, citation, referencing (like `\section`, `\cite`, `\ref`) and standard LaTeX environments (`theorem`-like environments, `enumerate` or `itemize` environments...) wherever possible. More details, template, can be found on our web site.

Marking of non-LaTeX papers. Greek, ghotic, or script letters, rare mathematical or other symbols should be identified or even explained if necessary in the margin. Scalar quantities and variables are to be denoted by italics (underline if you cannot typeset), vectors by lower case bold face letters (underline doubled), matrices by bold capitals (underline doubled). Dimensions (mm, Ohm,...), standard functions and operators (sin, ln,...) are to be typeset by upward Roman typefaces (not in italics). Emphasized words can be distinguished by italic setting (underline).

Graphics. The authors are encouraged to submit all figures also in electronic form. Figures created by a graphics package should be saved in encapsulated PostScript (EPS) format. If the graphics package used does not support EPS output, the graphics file should be saved in one of the standard graphics formats (JPEG,...). A separate file for each graphics should be submitted. EPS files may be incorporated into LaTeX file by using the standard LaTeX `graphicx` macro package.

Bibliography. For the style of the reference list, which is to be given in alphabetical order, see the examples below for journal articles, conference papers, and books. BibTeX users may use any of the standard `plain.bst`, `amsplain.bst` or our `pp.bst` BibTeX style file.

[1] EARL, A. – KIS, I. – TÖRÖK, I., Partial Discharge Measurement in Cables, *Per. Pol. Elec. Eng.*, **32** (1988), pp. 133–138.

[2] KISS, S. – SMALL, A. B., *Resistances*, Panem, Budapest, second ed., 1998.

[3] KISS, T., Roundoff Errors in FFT, *Proc. 15th IEEE Symposium on Signal Processing*, S. Smith ed., York, May 1996, IEEE Press, New York, pp. 133–139.

Copyright. It is the policy of Periodica Polytechnica to retain the copyright of the papers in order to protect the interest of its authors, and to facilitate the appropriate reuse of the material by others. At the same time Periodica Polytechnica doesn't restrict the author in the redistribution of his/her own work if the exact bibliographic data are attached to it. This means that the author may post the material on his/her own server, may use the material in education, and may send the work to preprint servers or databases, if it is clearly signed that the paper is submitted to/accepted by/published and copyrighted by Periodica Polytechnica.

CONTENTS

GÁSPÁR, P.–SZÁSZI, I.–BOKOR, J. : Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ Control Design for Active Suspension Structures 3

TARNAI, G.–SÁGHI, B.–MOSÓCZI, L.: The Conditions of the Application of ETCS in Hungary 17

ZUMKELLER, D. : Auswirkungen von Telekommuniktation und Verkehr auf räumliches Verhalten 23

JANOTA, A.: Using Z Specification for Railway Interlocking Safety 39

FRANK, J.–GYÖRGY, A.: Adoption of European Union's Law in Hungary, in the Area of Product Adequacy 55

MÉSZÁROS, P.: Towards Sustainable Mobility Conflicts, Impacts, Principles, Conditions, Aspects of CEEC 65

ELEÖD, A.–DEVECZ, J.–BALOGH, T.: Numerical Modelling of the Mechanical Process of Particle Detachment by Finite Element Method 77

TÁNCZOS, K.–RÓNAI, P. : Budapest – on the Way to Join the IDIOMA Project 91

TÁNCZOS, K.–FARKAS, GY.: Problems and Issues of Defining User Charges for Railway Infrastructure 103

TÁNCZOS, K.–KISS, B.: Ideas Reflected in the Future Public Transport Association in the Area of Budapest ... 117

KISS, G. K.: The Use of Modern Satellite Systems in the Railway Traffic 123

KULCSÁR, B.: LQG/LTR Controller Design for an Aircraft Model 131

WANG, X.–TAKÁCS, J.–KRÁLLICS, GY.–SZILÁGYI, J.–MARKOVITS, T.: Experimental Research on Laser-Material Interaction 143