Development of a new wear test method for hot forming

F. Tancsics[3], L. Solecki[2], E. Halbritter[1]

[1] Department of Materials Science and Technology, Széchenyi István University,
[2] Department of Vehicle Manufacturing, Széchenyi István University,
9026 Győr, Egyetem tér 1.
solecki@sze.hu, halbritt@sze.hu
[3] RÁBA Axle Ltd., Production Development,
9027 Győr, Martin út 1.
ferenc.tancsics@raba.hu

Abstract: Utilization of the lifetime extension possibilities of forming dies is one of the key questions in the economic production of forged products. Failure of the shaping surfaces is mostly due to wear processes (in about 70% of the cases) [1]. In order to calculate wear and to check experimentally the calculations the upsetting technology performed between parallel pressing plates has been chosen, which is used at the forging factory of Rába Axle Ltd. as a pre-upsetting (scale removing) step using robot technology. Local wear depth has been calculated by the Archard wear model. In order to apply the wear model one has to know the displacement field at the contact of the part and the pressing plate, the temporary pressure distribution and the wear coefficient characteristic of the die, depending on the working temperature of the die. In order to define the inter-dependent displacement fields the material flow has been mathematically modeled. Developing further the selected mathematical model, based on the largest diameter of the barreling part the approximate value of the friction coefficient has been determined, which is necessary to define the temporary displacement field and the temporary pressure values. We have also attempted to determine the wear coefficient experimentally. When evaluating the experiment special macro- and micro-geometrical tests were used. In order to solve the mathematical model numerically a program was written using the Mathcad software.

Keywords: material flow, wear, upsetting, die surface, friction coefficient

1. Cost analysis, actuality of the problem

Increase of the production costs of forged parts continued dramatically in the past few years (Fig. 1.). The reason of this change is partly the increase of the technical requirements and partly increasing material costs. A possible, obvious way of cost reduction is the increase of the lifetime of the dies by reducing the degree of wear being the main cause of die failures [2].

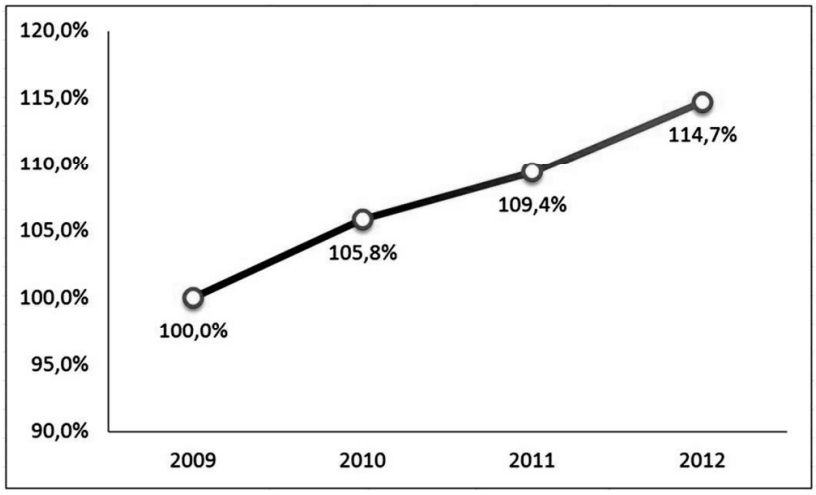


Figure 1. Changes in the specific production cost of forged parts (cost/kg, percentage change)

In multi-cavity impression-die forging it is an important requirement that the dies of the consecutive forming steps work in harmony and the die wear should be predictable. The study of die wear was therefore started at the upsetting dies characteristic of the first forming step of the process. In our earlier work a wear coefficient was determined for NK2 die steels by analyzing the abrasion marks of 26000 forged specimens prepared under the conditions used by Rába Axle Ltd. [3] [4].

Our aim in the present work is to investigate the possibility of joint determination of wear and friction coefficients and the more precise determination of the wear coefficient characterizing the die steels used.

2. Novelty of the method, used theoretical models

The practical modeling of the wear process of die surfaces with acceptable precision is made possible by the robot technology used in hot upsetting as it places the pieces to be formed always in identical position, according to the program (Fig. 2.).

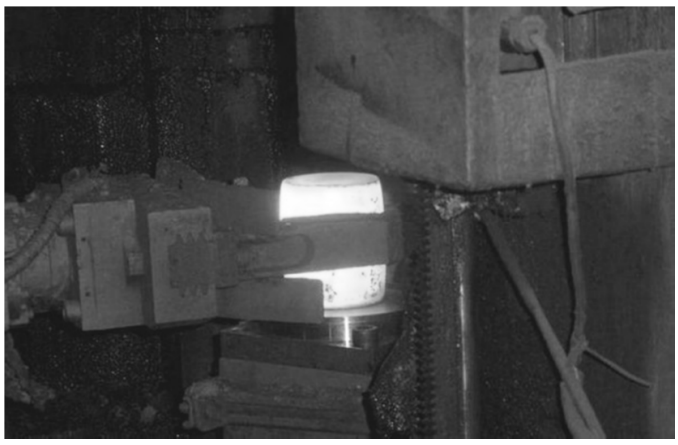


Figure 2. Manipulation of the upset forged part by a robot

In the modeling of the forging process the constant parameters are as follows: the surface roughness of the die, operation parameters of the die (scaled surface of the part, lubrication, kinematic and dynamic properties of the production machine etc.).

Based on all above the more precise determination of the wear coefficient is possible if the size of the experimental sample is reduced to the experimentally testable minimum. Under testable minimum that minimum is meant where the abrasion mark can be evaluated, the maximum wear site begins to be formed and the effect of adhesion wear is negligible. Then, at the beginning of the wear process the relative displacement of the contacting surfaces is influenced only by the friction coefficient, i.e. the abrasive wear produces the maximum mark.

Using this assumption the wear coefficient of the mark and the friction coefficient can be brought into direct relation with each other can be involved into an algorithm. This latter means that based on the average value of the maximum diameter of the upset forged part the friction coefficient and through this the wear coefficient of the die steel can be determined.

2.1. Mathematical modeling of material flow

Basic laws describing the motion of continua can be used for the mathematical modeling of material flow [5]. Such a basic law is conservation of mass:

$$\frac{\partial \varsigma}{\partial t} + \nabla \cdot (\varsigma \vec{v}) = 0 \tag{1}$$

where:

ς is the density of the material (kg/m^3),
t is the time (sec),
∇ is the Hamilton operator,
$\varsigma\vec{v}$ is the mass current (kg/m^{2}sec).

If considering plastic forming it is usually assumed that the density of the formed part does not change during upsetting, so form equation (1) the constancy of the volume described by equation (2) follows [5]:

$$div(\vec{v}) = 0\,. \tag{2}$$

When upsetting a primary part of cylindrical shape the deformation can be well approximated by axial symmetry, therefore the constancy of the volume may be expressed in the cylindrical coordinate system too, only / w_z / and / w_r / velocity components should be used:

$$div(\vec{w}) = \frac{\partial w_r}{\partial r} + \frac{w_r}{r} + \frac{\partial w_z}{\partial z} = 0 \tag{3}$$

where:

$\vec{w}$ is the velocity vector a given (r, z) point,
w_z is the axial velocity component at a given point,
w_r is the radial velocity component at a given point.

According to reference [6] the velocity field is kinematically allowed if within the body it satisfies everywhere the condition of incompressibility $\dot{\varepsilon}_{ii} = 0$ and the

circumferential boundary conditions. The boundary conditions that can be defined for the points of the processed part based on the movement of the pressing plates – in comparison to the friction-free case [7] [8] – should be complemented with the assumption that / w_z / the axial velocity component (4) has an inflection point at z=h/2, i.e. at this point the deformation rate / $\dot{\varepsilon}_z$ / exhibits an extremum. The model with the assumptions listed above results in a barreled part. The axial velocity component of the kinematically allowed velocity field can be described by a third order polynomial:

$$w_z(z) = az^3 + bz^2 + cz + d\,. \tag{4}$$

The velocity field of the points of the barreling part (Fig. 3.) are described by the following functions:

$$w_i(r,z) = [w_r(r,z); w_z(z)] \tag{5}$$

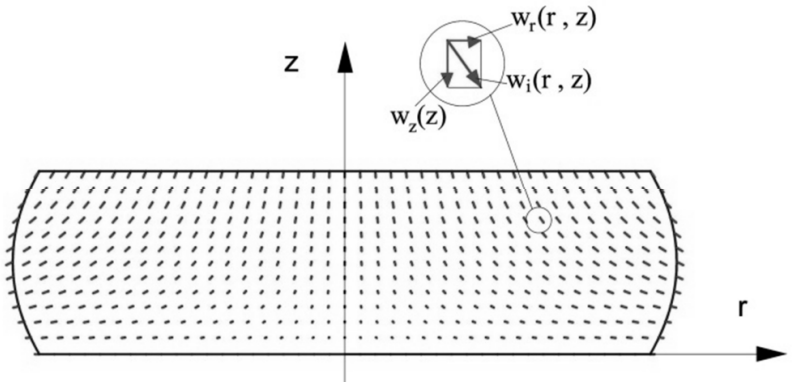


Figure 3. Image of the velocity field resulting in bulging part

Modifying parameters /c/ of equation (4) to a dimensionless parameter k=-ch/v_0 the velocity functions (taking into account the boundary conditions) can be reproduced as equations (6) and (7) [7] [8].

The axial velocity component is:

$$w_z(z) = \frac{z\left(-2z^2kv_0 + 2z^2v_0 + 3zkv_0h - 3zv_0h - kv_0h^2\right)}{h^3}. \tag{6}$$

The radial component of the velocity field / w_r (r,z) / can be obtained by solving equation (3):

$$w_r(r,z) = -\frac{1}{2}\frac{r\left(-6z^2kv_0 + 6z^2v_0 + 6zkv_0h - 6zv_0h - kv_0h^2\right)}{h^3}. \tag{7}$$

The exact value of / k / can be determined by minimizing the power-demand of the forming process [7] [8]. If using the assumed velocity field this dimensionless parameter influences the degree of barreling which, in turn is related to the friction coefficient. Therefore the value of / k / is related to the Kudo friction coefficient [7]. The relation between the two factors can be most easily given by equation (8) [7] [8]:

$$m = 1 - k\,. \tag{8}$$

In the case of a pre-upsetting task the following parameters are given: the initial radius / R_0 /, the initial height / H_0 / and the height of the upset part / h_n /. Using these data and equations (6) and (7) one can simulate the profile curves, the / R_{min} / and / R_{max} / values as a function of / m / (Fig. 4.).

At the applied simulation the value of initial height / H_0 / was cut down with v_0dt=0.1 mm, and the new position of the points was calculated by using the relations (6, 7). Then this new geometry was considered to be the initial one, and the previous steps were repeated by decreasing the heights again. This cycle was repeated until the height at an instant reached the specified value of / h_n /. The stabilities of / k / and / m / are rightfully assumable in case of summing up elemental vertical shifts of upsetting. We have presumed the stability of these values throughout small pre-upsetting.

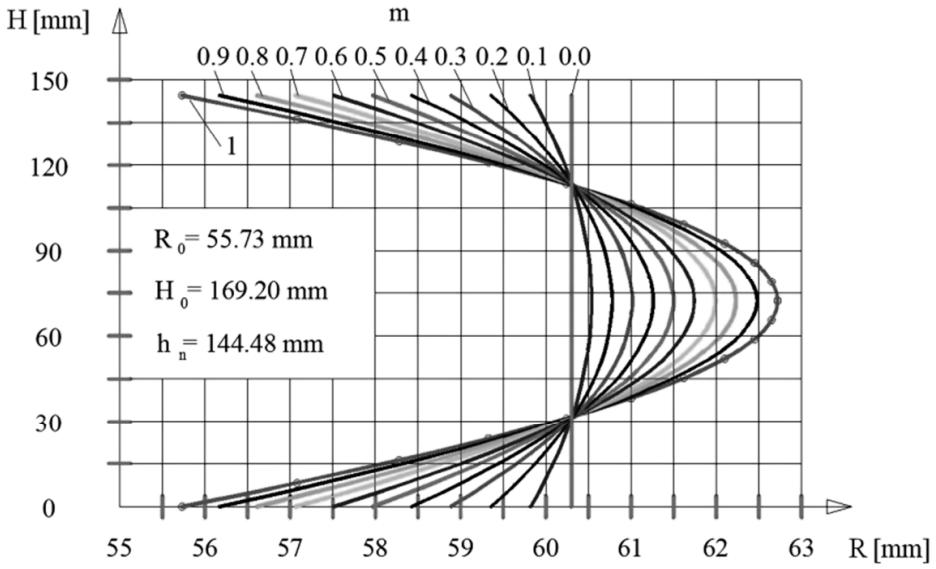


Figure 4. Profile curves as a function of the Kudo friction coefficient, m

Specific temperature conditions of the forming process should be taken into account in the evaluation process. During upsetting at the contact point of the formed part and the pressing plate the temperature of the part (with some simplification) is close to 1100 °C, that of the pressing plate to 300 °C.

The corrected radius of the part is calculated using linear thermal expansion:

$$R_{1100} = R_{20}\left(1 + \alpha_{pt}\Delta T'\right) \tag{9}$$

where:

α_{pt} is the linear thermal expansion coefficient of the part: $12x10^{-6}$ ($°C^{-1}$),
R_{1100}= 55.726 mm.

Of course other geometrical parameters of the part / H_0, h_n / should also be modified. The simulated wear results thus can be related to compression plates of 300 °C temperature, so data for room temperature can be obtained by further calculations. The thermal expansion of the diameter of the die contacting the part is:

$$\Delta R_{300} = R_{20}\alpha_{sz}\Delta T \tag{10}$$

where:

α_{SZ} is the linear thermal expansion coefficient of the die: 10.37×10^{-6} ($°C^{-1}$),
$\Delta R_{300} = 0.175$ mm.

There is a functional relation between radii / R_{min} / and / R_{max} / on the one hand and the / m / coefficient on the other, and this relation can be fitted by a regression curve (a second order polynomial) for the given upsetting operation (Fig. 5.).

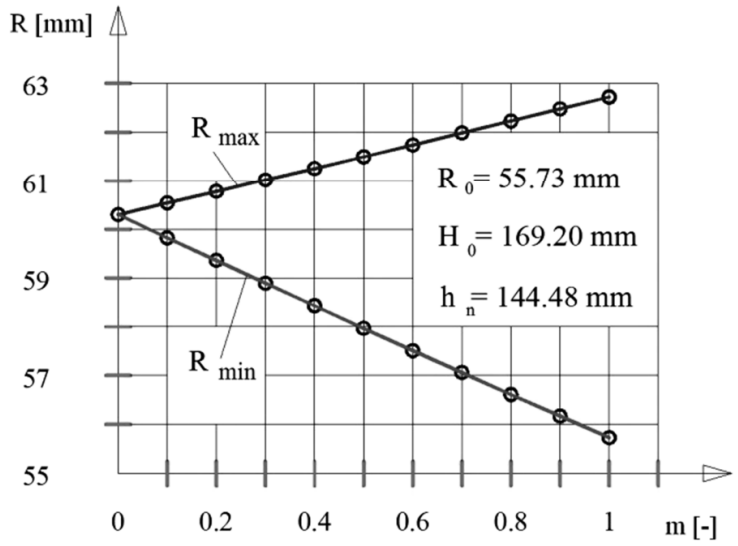


Figure 5. R_{min} and R_{max} values plotted against the Kudo friction coefficient, m

The inverse functions $m=f(R_{min})$, or $m=f(R_{max})$ can also be written, from which the friction coefficient can be determined for a given / R_{max} / radius.

In order to use this method it had to be proven that the profile curves obtained by modeling the material flow correspond well to the possible profile curves of real parts[8] [9] [11]. When proving the relation first we demonstrated using literature data and own test results that – within the proposed geometrical constraints [12] - the profile curves can be well fitted by a second order polynomial [7] [8] [10].

The second order polynomial describing the profile curve of real parts (equation (11)) can be defined by the minimum and maximum radii / R_{min} and R_{max} / and the temporary height of the upset part / h /.

$$R(z) = -\frac{4(R_{max} - R_{min})}{h^2} z^2 + \frac{4(R_{max} - R_{min})}{h} z + R_{min}. \qquad (11)$$

The determination of the minimum radius / R_{min} / in practice is problematic, as the barreled outer surface (mantle) may join other surfaces not only along an edge, but also with a small radius. The reason of this that during the real shape evolution process sticking of various degree may also appear, resulting in a curvature of various radius (experimentally measured value is about 2.5 mm) between the barreled mantle and the flat face. Equation (11) is valid without taking into account the curvature mentioned above, i.e. the mathematical model does not take into account sticking the friction coefficient obtained should be regarded as an approximate, average value [8]. Based on the assumption of constant volume one can write:

$$R_0^2 \pi H_0 = \int_0^h R^2(z)\pi dz \cdot \tag{12}$$

After inserting equation (11) one can express / R_{min} / from equation (12), i.e. the profile curve can be drawn if the initial radius / R_0 /, the initial height / H_0 /, the upset height / h_n /, and the maximum radius of the upset part / R_{max} / are known, a high degree of similarity (and inter-relation) between the profile curves obtained from the modeling of material flow and those possible geometrically can be proven.

The friction coefficient value can be determined from the given initial radius / R_0 /, initial height / H_0 /, upset height / h_n / and from the maximum radius of the upset part / R_{max} / [15].

2.2. Mathematical modeling of the wear process

One of the best known relations describing abrasive wear is the Archard wear model which (as a function of the wear coefficient characterizing the die) assumes proportionality between the contact pressure of the contacting surfaces, the displacement and the worn out material volume (13):

$$dV = K(T)\frac{dF_n}{H(T)}dL \tag{13}$$

where:

dV is the volume of the worn out material (mm^3),
K(T) is the wear coefficient characterizing the die, depending on the working temperature of the die (-),
dF_n is the normal force acting on the contacting surfaces (MN),
H(T) is the Brinell surface hardness (HBS) of the die depending on the working temperature of the die (MPa),
dL is the displacement of the contacting surface elements (mm).

Simplifying assumptions:

- the temperature of the upsetting die does not exceed 450°C during upsetting, i.e. its hardness can be regarded as constant [13],
- the part to be upset is characterized by a homogenous temperature field,
- the temperature and the working strength during upsetting are homogenous and constant,
- only the lower die is investigated because of the higher heat load,
- there is no sticking during the study,
- the friction coefficient values in the study on both die-halves are considered to be constant and identical in course of the upsetting process.

Taking into account the simplifying assumptions, based on the Archard wear model the wear depth for one upsetting cycle can be determined numerically [1]:

$$z = K\sum_{i=1}^{n}\frac{\sigma_n(r,t)_k\, v(r,t)_k}{H}\Delta t \tag{14}$$

where:

z is the wear depth (mm),
K is the specific wear coefficient characterizing the die (-),
n is the number of the discretized increments of the height reduction (pc),
$\sigma_{n(k)}$ is the normal stress at the contacting surface elements in the /k^{th}/ increment (MPa),
$v_{(k)}$ is the relative slipping velocity of the contacting surface elements in the /k^{th}/ increment (mms^{-1}),
H is the Brinell surface hardness (HBS) of the die (MPa),
Δt is the contact time interval of the surface elements, the time increment during the displacement (sec).

When modeling the upsetting cycle by numerical mathematical methods the forming process within the upsetting cycle should be divided into /n/ subsequent intervals (steps). Within one step the upper pressing plate moves $\Delta h=v_0\Delta t=0.1$ mm. At every new step a new height value / h / should be used. The actual height is obtained if the / Δh / value is subtracted from the previous height value. The functional relation at the contact of the pressing plates and the upset part can be written as follows, using equation (8):

$$w_r(r,0)=w_r(r,h)=\frac{1}{2}\frac{r(1-m)v_0}{h}. \tag{15}$$

When studying wear a displacement-field can be defined instead of the velocity field. Equation (5) changes accordingly to:

$$u_r(r,0)=u_r(r,h)=\frac{1}{2}\frac{r(1-m)\Delta h}{h} \tag{16}$$

where:

Δh is the displacement of the upper pressing plate during the / Δt / interval $\Delta h=0.1$ (mm).

It can be well seen from equation (16) that the displacement value increases linearly along the radius, and its value is zero at r=0 (Fig. 6.). At r=0 there is no displacement, the die is not worn. This is, of course, only a theoretical statement, valid for point-like surface of infinitesimal size, but this train of thought should be considered when assessing the expected wear distribution.

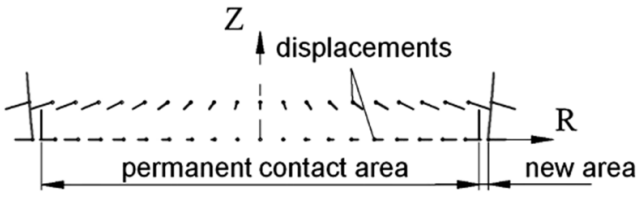


Figure 6. Division of the contact area

The friction coefficient value influences the radial displacement, thus the wear. If the upsetting process would be frictionless, m=0, the part would not barrel and the relative displacement would be at maximum. In case of complete sticking m=1 the surface of

the part would not change when contacting the die. From the viewpoint of wear the displacements should be summed up step by step. During the summation process the initial area under the ingot should be treated separately from the contact surface parts formed in course of the upsetting process. In the Archard wear model (see equation (14)) the stepwise displacements and the actual pressure should be considered together. In the upsetting process the pressure arising at the contact of the parallel pressing plates and the formed part is not uniform. According to literature hints [14] when upsetting solid cylindrical parts by parallel pressing plates the surface pressure / p / can be calculated form the working strength / k_f /, the Coulomb friction coefficient / μ /, and geometrical data / h, R, r / as follows:

$$p(r) = k_f e^{\frac{2\mu}{h}(R-r)} \tag{17}$$

The exponential equation (17) can be simplified by Taylor expansion and by neglecting higher order members. There might be a more accurate solution [16] but this approach is easy to use for MathCAD programs and gives correspondent and accurate results for practice. An approximate relation for the surface pressure taking into account the Kudo friction coefficient is as follows:

$$p(r) = k_f \left(1 + \frac{2m}{\sqrt{3}h}(R-r) \right) \tag{18}$$

The simultaneous consideration of displacements and pressures is easier by the numerical method. Wear depth at r=0 is zero, as there is no displacement, but, at the same time the surface pressure is maximum at this point.

3. Industrial experiment

In order to prove the theory industrial experiments were performed. Before starting the experiment a silicone replica was taken of the active surface of the upsetting dies (Fig. 7.). The average surface roughness of the upsetting dies were machined to Ra=0.25 in order to shorten the adhesive wear process and thus to minimize the sample quantity. Useful orientation points were machined onto the surface to support the evaluation (see the red arrows).

Figure 7. Taking silicone replica of the active surface of the lower upsetting die

The industrial experiments were made under constant production conditions in the forging factory of Rába Axle Ltd. A picture of some of the experimental specimens is shown in Fig. 8.

Some characteristic data are as follows:

- cut mass: m'=12.48 ±0.2 kg,
- cut length: H_0=167 mm,
- heating temperature: T_{heat}=1213-1226 °C → $T_{(average)}$ =1219 °C,
- initial diameter: D_0=110 ±0.2 mm,
- upset height: $h_{n(coolsize)}$=142.32-142.91 mm → $h_{n(average)}$ =142.6 mm,
- upset upper diameter: $D_{1(upper\ -\ coolsize)}$=114,2-114,5 mm → $D_{1(average)}$=114,3 mm,
- upset lower diameter: $D_{2(lower\ -\ coolsize)}$=112,4-113,7 mm → $D_{2(average)}$=113,0 mm,
- upset largest diameter: $D_{k(coolsize)}$=121,6-122,3 mm → $D_{k(average)}$=122,1 mm,
- lubrication: without lubrication,
- forming equipment:10 MN LASCO hydraulic press,
- hardness of die surface: 48 ±2 HRC,
- measured die surface temperature: $T_{lower-max.}$=302 °C; $T_{upper-max.}$=239 °C.

Figure 8. Forged parts of the upsetting experiment

The surface temperature of the dies did not reach the critical threshold value of 450 °C where softening starts [13].

4. Simulation of the wear process

Using the numerical relations introduced above an own Mathcad program was developed [3] to predict the integral displacements and the expected wear characteristics (the location and degree of the largest wear). Based on the functional relations established earlier for the surface hardness and wear coefficients of NK2 die steels used in Rába Axle Ltd. [3], the wear coefficient was chosen as $K=5.49\times10^{-5}$. Using the program the initial value and location of the expected dry wear in one upsetting cycle and the expected size of the worn surface were determined. Input data were: H_0=169.20 mm; R_0=55.73 mm; h_n=144.48 mm; m=0.7; k_f =100 MPa; $K=5.49\times10^{-5}$.

When defining the friction coefficient, as an initial value, the average of the maximum upset diameters of the specimens measured in cool state were taken into account: $D_{k(average)}$=122.1 mm. Based on equation (9) the maximum upset radii expected at the forming temperature was also determined: $R_{max(1100)}$=61.86 mm. The Kudo friction coefficient can be directly read from Fig. 5, at the radius of R_{max}=62 it is m=0.7. The friction coefficient necessary for the simulation can also be obtained from Fig. 4. In

this case the radius of R=62 identifies the profile curve belonging to the friction coefficient of m=0.7. Using equation (14) the program calculates with the input data the expected largest extension of the mark, its depth and location under the simplifying assumption borne out by the experiments that the surface hardness of the die can be regarded as constant during the upsetting process [3] [13]. The results of the run are shown in Fig. 9.

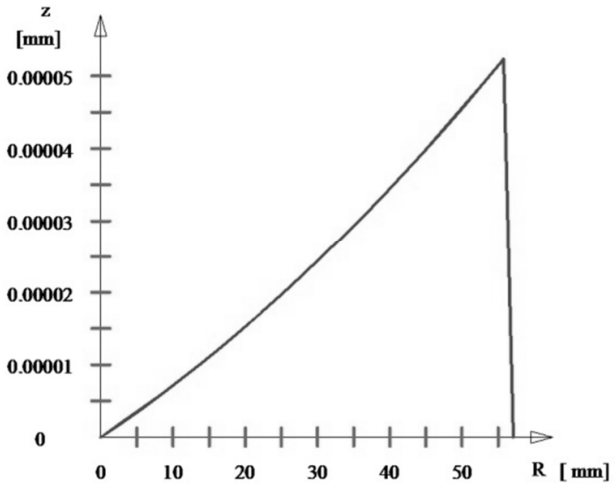


Figure 9. Wear distribution for a single upsetting cycle

Before the test the expected diameter of the abrasion mark was estimated based on the size changes due to the different thermal expansion coefficients of the contacting surfaces of the part and the die. The expected diameter of the mark at room temperature is estimated using equations (9) and (10): 114.49-0.35=114.14 mm. From Fig 9 the expected maximum wear depth is 5.2293×10^{-5} mm for a single upsetting cycle, its expected distance from the center is 55.55 mm using equations (9) and (10). The expected maximum radius of the abrasion mark is: 57.06025 mm which corresponds to a diameter of 114.1205 mm. From the viewpoint of forging the decisive factor is the surface element exposed to maximum wear, which is sensitive to the material flow therefore it is a potentially dangerous site for surface folds on the formed part. The results obtained were also projected to 26000 workpieces [3]. The calculated maximum wear depth was: 1.3596 mm. In case of 26000 workpieces it is already necessary to take the diffusion heat transmission processes of tool surface into consideration. These processes cause the soft layer on the surface. Therefore the depth of real wear rate may be slightly higher.

The Mathcad program yields only approximate wear data but its great advantage is that it can be easily joint with CAD systems and the 3D geometry necessary for design can be parametrically defined. Afterwards the largest extension of the mark was investigated by various test techniques.

5. Comparison of the macro- and micro-geometries of the new and worn pressing plates

After upsetting the experimental parts (Fig. 9.) the lower pressing plate and the upset part No. 66 were investigated carefully in the laboratory of the Széchenyi István University. The goal of this study was mainly to determine the maximum extension of

the abrasion mark and to detect the changes within the contact area between the part and the pressing plate.

5.1. Macro-geometry of the abrasion mark

Within the contact area the surface roughness of the pressing plate changes, sooner or later worn-out cavities are formed and, due to the heat effect, the pressing plate became discolored.

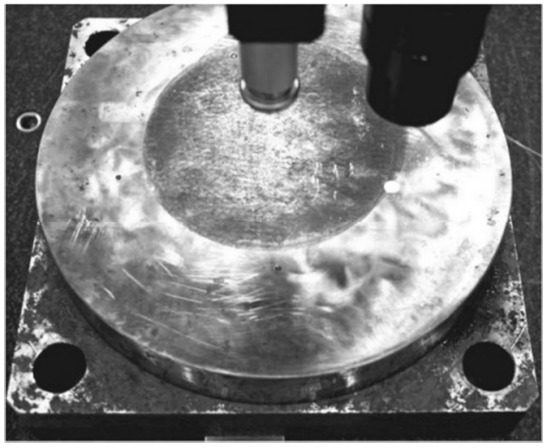


Figure 10. Monitoring the borderline of the contact area with a MAHR PMC800 coordinate tester

The theoretical division of the contact area was shown in Fig. 6. The permanent contact area and the newly formed area can be experimentally observed (Fig. 10.). The borderline between the permanent contact area and the newly formed area can be approximated by concentric circles (Fig. 10.). Deviations from the circular shape can be due to the anisotropic properties of the material, with the macro- and micro-geometry of the contacting surfaces of the pressing plate and the formed part before upsetting and with positioning uncertainties of the part.

The surface of the pressing plate was ground after fine milling. Milling resulted in an ordered pattern, grinding in a disordered one. Grinding was uneven sometimes it left in patches the original milling pattern on the surface (Fig. 11.).

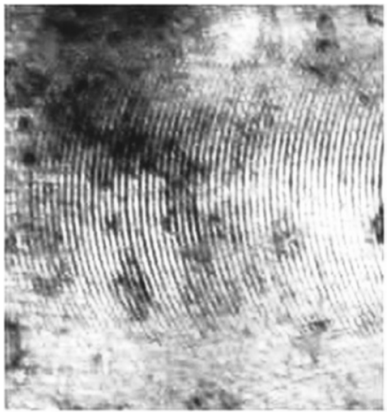


Figure 11. Local residues of the milling pattern on the pressing plate

The initial parts for upsetting were sawn from a rod. After upsetting the sawing pattern can be unambiguously observed on the pressing plate. The sawing pattern is

mostly caused by discoloration due to thermal overload, it can be observed visually, but cannot be detected at a macro-geometrical level. This pattern partly survived on the face of the upset part (Fig. 12.).

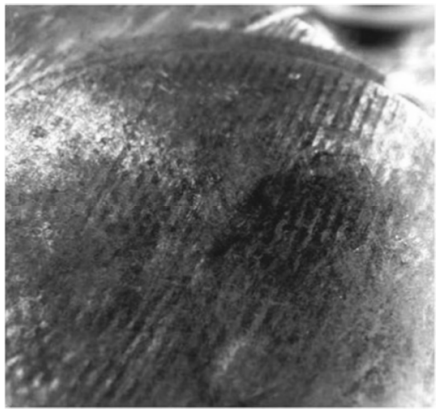


Figure 12. Sawing pattern from the end-plate of the formed part on the pressing plate

When studying abrasive cavitation the comparison of the micro- and macro-geometries of the pressing plate before and after use was performed on silicone replicas using a Taylor – Hobson Talysurf CLI 2000 roughness tester. Using the 4 orientation (reference) points it was possible to position the scanning of the worn surface within 0.1 mm to the original one. Scanning was done in two, mutually perpendicular directions.

It has been established that the silicone replicas could not be used for macro-geometrical comparison, as the contacting surfaces of the replicas were larger and less accurate than the stage of the roughness tester, so the flaw and error-free positioning of the silicone replicas was not possible.

Using a MAHR PMC800 coordinate tester the limiting points of the abrasion tracks were determined and the approximate diameter of the circle was determined (Fig. 13.).

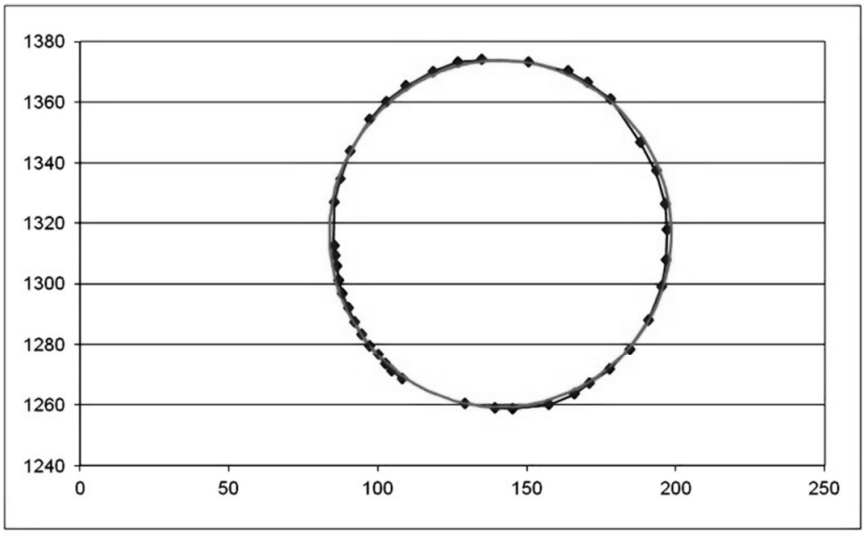


Figure 13. Coordinates of the center of the D_2 circle approximating the data: y=1316.508, x=141.164

The number and position of the points selected for fitting the approximating circle influence the end result, so the arithmetical mean of the diameters of two approximating circles with different numbers of points (D_1 and D_2) were taken into account. The arithmetical mean was: $D_A=(D_1+D_2)/2=(113.489+114.670)/2=114.0795$ mm.

The diameter of the approximating circle can also be determined by optical analysis using the photograph of the mark. The advantage of this approach is that using a proper CAD system – in our case AutoCAD – the 3D position of the points is reduced to 2D and the points used for the evaluation can be positioned at freely selected magnification. Effective processing of the evaluation points can be done by proper program development, by a joint use of AutoCAD and Mathcad software. Using AutoCAD the diameter of the approximating circle is also: D=114.08 mm (Fig. 14.).

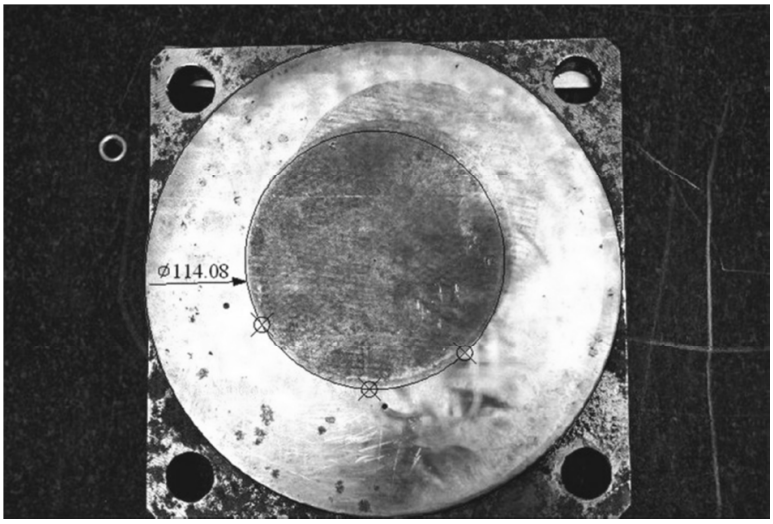


Figure 14. Determination of the approximating circle by the AutoCAD program

The remarkable resemblance of the experimentally determined parameters, in spite of the weak points of the different methods used can be regarded as an important finding. The test results can be well compared with the simulation results of our Mathcad program, the program can be corrected. The negligible amount of correction (0.14 mm, the rounding of $R_{max}=61.86$ to $R_{max}=62$) corroborates the applicability of Fig. 5 describing the relation between the friction coefficient and the abrasion mark.

5.2. Micro-geometry of the abrasion mark

In order to detect changes in the micro-geometry 2x9 sections of 5 mm length were scanned for surface roughness determination, as shown in Fig 15. Ra and Rz roughness values were evaluated after removing shape error and the waviness was removed by a 0.8 mm Gauss filter.

Sections 1 and 9 were on the reference are outside the abrasion mark, where the part did not contact the pressing plate.

Sections 2-8 were situated within the abrasive mark in equidistant positions, section 5 was at the center.

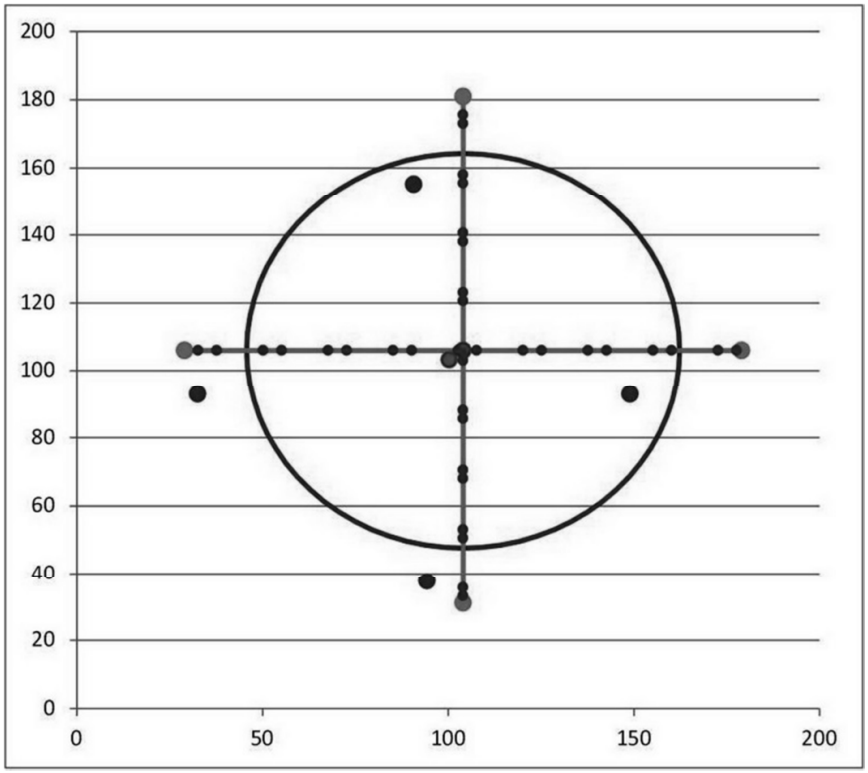


Figure 15. Sections evaluated for surface roughness

Table 1 shows the results of surface roughness evaluation for the various sections.

Table 1

row	original (x)		worn (x)		original (y)		worn (y)	
	Ra	Rz	Ra	Rz	Ra	Rz	Ra	Rz
1	0.292	2.28	0.418	3.71	0.200	1.90	0.364	4.01
2	0.311	2.87	0.590	6.53	0.210	2.03	0.593	5.96
3	0.265	2.51	0.752	1.01	0.215	2.69	0.508	4.37
4	0.313	3.38	0.468	3.66	0.240	2.20	0.415	3.94
5	0.360	3.00	0.526	4.96	0.320	3.05	0.501	4.60
6	0.288	2.59	0.412	3.18	0.462	3.74	0.653	4.58
7	0.328	5.02	0.579	6.16	0.344	3.20	0.502	4.43
8	0.858	5.49	1.290	7.45	0.513	5.21	0.557	4.53
9	0.655	4.87	0.852	6.13	0.482	5.92	0.597	7.05

It can be concluded that the whole surface of the pressing plate became rougher and the roughness value at the edge of the abrasion marks– where the expected wear is maximal – is much larger than at other places (Fig. 16.).

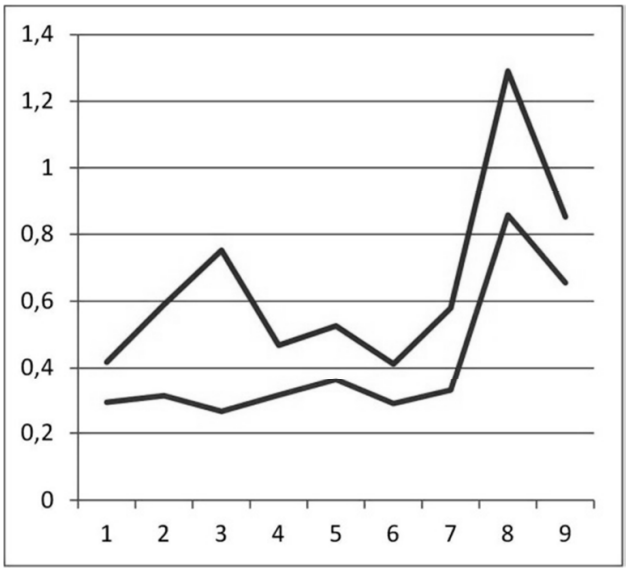


Figure 16. Average roughness (Ra) at various sections in the x direction (lower curve - initial, upper curve – worn surface)

Those sites can be unambiguously identified where polishing did not remove milling patterns (Figures 17. and 18.).

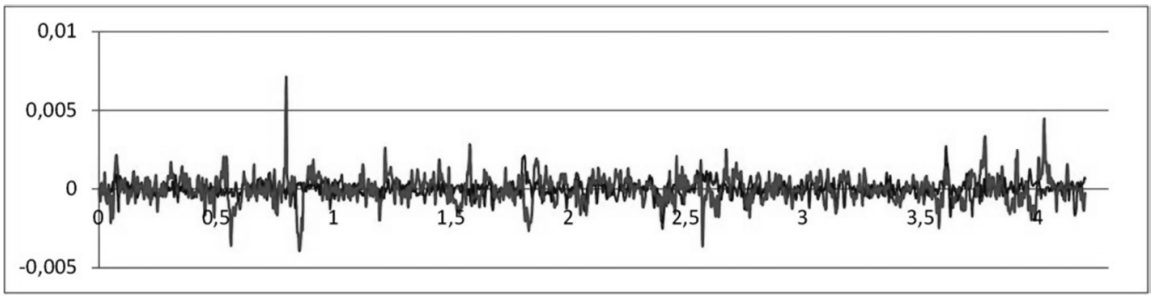


Figure 17. Surface roughness profile at an area without residual milling pattern (lower amplitude) and in worn state (higher amplitude) (axes are given in mm units)

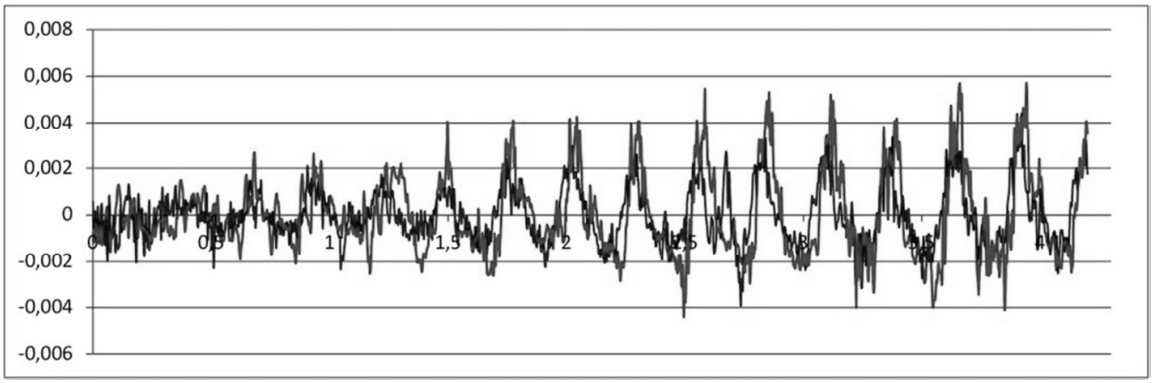


Figure 18. Surface roughness profile at an area with residual milling pattern (lower amplitude) and in worn state (higher amplitude) (axes are given in mm units)

During upsetting the pressing plate is loaded by an uneven compressional load. The stress distribution corresponds to the pressure distribution given by equations (17) and (18). Under the effect of the arising compressional stresses the flat surface of the pressing plate is deformed elastically. When upsetting the part – with some

simplification – it is deformed only in a plastic manner and the face of the formed part inherits the elastic deformation of the pressing plate.

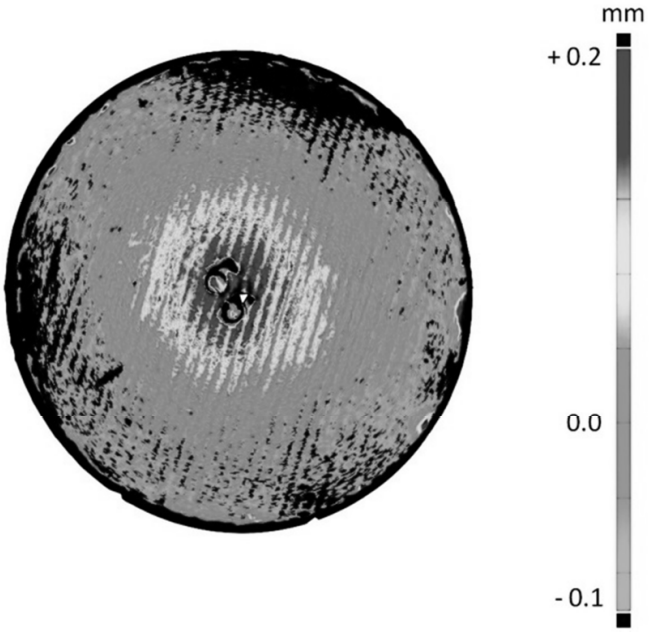


Figure 19. A protrusion observed on the face of the part to be formed

Deviations form planarity were studied separately using the MAHR PMC800 coordinate tester. Scanning the studied surface systematically 216000 data were collected. The test data were analyzed as a point cloud using the GOM evaluation software.

It was observed that the central part of the face plate protrudes and the sawing pattern can be partly observed (Fig. 19.). In our opinion the residual sawing pattern may be due to non-removed scale on the face of the formed part. The protruding part can be well approximated by a spherical surface with a radius of R=4606 mm. The protrusion is so small (0.35 mm) and structured (Fig. 19.), that in order to detect it one must use a coordinate tester with a precision of 0.001 mm.

6. Conclusions and Future Improvements

In our earlier work we suggested a method to determine approximately the friction coefficient. In our present work we proved that the proposed method can be used under real production conditions - practically without any additional cost – for pre-upsetting (scale removal). Pre-upsetting on robot assisted lines made possible the joint investigation of the friction coefficient and wear.

We have proved by a method, which is new in technology planning that the wear coefficient can be related to the friction coefficient and the method can be well algorithmized. A more precise determination of the wear coefficient requires further studies. We would like to investigate the possibility of taking into account the frictional work.

References

[1] Zamani, A.M., Biglari, F.R.: *Finite-element investigation of wear in hot forging*, Tehran International Congress on Manufacturing Engineering (TICME2005), Tehran, (2005), pp. 1-9.

[2] Barrau, O., Boher, C., Gras, R., Rezai-Aria, F.: *Analysis of the friction and wear behaviour of hot work tool steel for forging*, 6th International Tooling Conference, Karlstad (2002), ©2003 Elsevier Science B.V. pp. 95-111.

[3] Tancsics, F., Halbritter, E.: *Melegalakító szerszámok kopásvizsgálata*, A Jövő Járműve 01/02, (2012), pp.29-35 (in Hungarian).

[4] Solecki, L., Halbritter, E.: *Macro-and Microgeometrical Comparison of the Surfaces of Forming Dies*, 13th International Conference on Tools (ICT-2012), Miskolc, (2012), pp. 245-250.

[5] De Arizon, J., Filippi, E., Barboza, J., D'Alvise, L.: *A Finite Element simulation of the hot forging process*, Service de G´enie M´ecaniquepp, (2006).

[6] Prager, W., Hodge, P.G.: *Theory of perfectly Plastic Solids*, Wiley, (1951)

[7] Halbritter, E.: *Modeling of Material Flow During Upsetting Between Parallel Pressure Plates*, Hungarian Electronic Journal of Sciences, (1999), ps. 11.

[8] Tancsics, F., Halbritter, E.: *A súrlódási tényező újszerű meghatározása és felhasználása a Pro/Engineer és MathCAD szoftverek segítségével*, GÉP LXI/7, (2010), pp. 34-42 (in Hungarian).

[9] Chen, Z.Y., Xu, S.Q., Dong, X.H.: *Deformation Behavior of AA6063 Aluminium Alloy after Removing Friction effect under hot Working Conditions*, Acta Metallurgica Sinica, Vol.21 No.6, (2008), pp. 451-458.

[10] Tancsics, F., Halbritter, E., Kiss B.: *Simplified Determination of Friction Coefficient by Upsetting*, OGÉT 2009 17th International Conference on Mechanical Engineering, Gyergyószentmiklós, (2009), pp. 384-387.

[11] Bartoň, S., Hřebíček, J.: *Heat Flow Problems, in Solving Problems in Scientific Computing using MAPLE and MATLAB.*, edited by Gander, W. & Hřebíček, J., Springer, (2004), pp. 191-200.

[12] Tancsics, F., Kiss, B., Halbritter, E.: *Limit Analysis of Adaptation of the Mathematical Model Made to Determine Friction Coefficient, OGÉT 2011 19th International Conference on Mechanical Engineering*, Csíksomlyó, (2011), pp. 355-359.

[13] Behrens, B.A., Schäfer, F., Hundertmark, A., Bouguecha, A: *Numerical analysis of tool failure in hot forging processes*, Obróbka Plastyczna Metali t. XIX nr 4, (2008), pp. 11-17.

[14] Siebel, E.: *Die Formgebung im bildsamen Zustand*, Verlag Stahleisen, (1932).

[15] Tancsics, F., Halbritter, E,: *Determination of Friction Coefficient During Upsetting Using a Kinematically Admissible Velocity Field,* Strojnícky Časopis, Journal of Mechanical Engineering, Vol.63 No.4, (2012), pp. 197-223.

[16] Ebrahimi, E., Najafizadeh, A.: *A New Method for Evaluation of Friction in Bulk Metal Forming*, Journal of Materials Processing Technology, 152, (2004), pp. 136–143.

Evaluation of railway track geometry stabilisation effect of geogrid layers under ballast on the basis of laboratory multi-level shear box tests

Ferenc Horvát, Szabolcs Fischer, Zoltán Major

Department of Transport Infrastructure and Municipal Engineering
Széchenyi István University
Egyetem tér 1. 9026 Győr, Hungary
Phone: +3696503400
e-mail: horvat@sze.hu, fischersz@sze.hu, majorz@sze.hu

Abstract: In this article authors investigated the railway track geometry stabilisation effect of geogrid layers under ballast with a specific laboratory multi-level shear box. During the laboratory tests four different types of geogrid layer (in two cases combined with geotextiles) were analysed when railway ballast was uncompacted and compacted. Two types from these (geogrid type 2 and geocomposite type 2) have not utilised for railway track geometry stabilisation yet. The authors determined inner shear resistance of railway ballast in case of without and with geogrid reinforcement, as well as five multiplication factors were defined which are adequate for determining inner shear resistance of reinforced and unreinforced railway ballast in consideration of different parameters.

Keywords: railway, track faults, geogrid-reinforced ballast

1. Introduction

1.1. General introduction remarks

Environment protection and awareness will be more and more important, therefore technologies should be used in all area of life which can conform considerably to this goal. It has to be noted that all the industrial and all transportation sector use fossil energy source, therefore quantity of CO_2 and other harmful materials increase in the atmosphere [9]. People of future should think about decreasing high rate of personal vehicle (automobile) transport in modal split or using more environmental-friendly energy sources because of their subsistence and ensuring of healthy life. Reversing of modal split to the adequate direction can be simply achieved with popularization of public transport, but service standard and high fares should be normalized. Against the road transport environmental friendly electric hauling railway transport has to be preferred which should be greater and greater part of continental traffic in case of passenger and freight transport too.

Rapid reaching of these aims significant financial support (foreign supports as well as national fund) has to be available, from which construction and reconstruction of high standard of railway infrastructure and investments of modern railway vehicles can be financed. If the author's scenario is fulfilled, economic execution of construction and rehabilitation reconstruction as well as decreasing of number of maintenance works will be important. In the author's opinion utilizing of modern building material and technologies will be indispensable.

1.2. Motivation and aims

It has to be noted that this paper only deals with conventional ballasted railway tracks and their more permanent geometry stabilisation. The reason of this fact that

- in the world 98.8% of railway lines is ballasted railway tracks (approximately 1.1 million km), only 1.2% is high speed railway slab tracks and maglevs [20],
- in Hungary there are slab track only on bridges and in tunnels,
- at maintenance works of slab tracks emphasis is totally different. In the consideration of speed of deterioration process of ballasted tracks, mainly respect to track geometry, they grant more disadvantageous solution.

In reference to railway infrastructure developments mentioned in Section 1.1 phrasing of more worry seems to be topical. In the past two decades[1] the Hungarian governments in power didn't pay enough money for the railway maintenance works, which would have been necessary. However, infrastructural developments were done from ISPA, KÖZOP, EU, EIB, PHARE, etc. funds, but their whole length to the Hungarian railway network is very low. Money, which can be spent for railway maintenance works, is very scant, so this is the reason of the fact that all the track faults can't be eliminated.

The extant track faults will be deteriorated forward due to the cancelled (or delayed) maintenance, and other new faults can be evolved[2] [6, 16]. If the size of the track fault exceeds the prescribed value contained the maintenance regulations related to railway tracks [10] speed restrictions have to be introduced [5], i.e. a reduced speed is allowed in this section. At the end of this kind of sections additional acceleration energy demand comes forward as compared to state if the train can be driven with constant speed.

Keep on this train of thought, it has to be mentioned that deterioration process of railway track is a natural physical procedure that is unstoppable and irreversible, with in time executed and professional maintenance work can only be slowed down. There should be an initial track fault and destructive effects for the deterioration process. There is track fault as a dimensional deviation, because engineering structures can't be constructed and maintained without any dimensional deviation. The destructive effect means primarily traffic, but environmental effects (e.g. weathering) can't be forgotten. Deterioration process of railway track can be divided in two different parts: geometrical deterioration period and structural destruction period [19]. Primarily geometrical

[1] Until 1992 adequate maintenance was ensured, early in 1993 approximately 1600 km speed restrictions should be introduced because of the drastically decreased financial supports.

[2] There are a lot of speed restriction segments in Hungarian railway lines, all one increases energy consumption of trains, as well as journey time is also heightened which worsens service standard of railway traffic and transport. **Appropriate technical-political verdicts should be returned.**

deterioration period has to be dealt with, in which distortion of railway track is evolved firstly as dimensional deviation, secondly as dimensional faults. If the speed of deterioration process should be slowed down, regulation work (e.g. tamping) is not enough, but structural change has to be made. This kind of change is e.g. using heavier rails, sleepers, more modern flexible and higher clamping force ensured rail fastening; reinforcement of superstructure with ballast gluing, or using of geosynthetics; substructure reinforcement, etc. In this way track faults can be evolved more slowly as well as time interval between regulation works can be lengthened, which has great financially importance of course[3]. In case of any track fault is there, introduction of speed restriction is unnecessary, therefore costs of additional acceleration energy due to speed restriction don't debit the operator of railway vehicles [5].

From the above mentioned structural changes the authors chose the geogrid layers under railway ballast, because this technology is used for a few years with little practical experience and appreciating analysis for railway track geometry stabilisation[4]. Figure 1 and 2 show interlocking effect of geogrids. The essence of increasing of inner shear resistance and strength of layer structure means acting together of geogrid layer and crushed granular material. The particles of crushed granular material are wedged into the aperture of geogrid and interlocked with the geogrid ribs. In this manner a quasi-strong and relatively skidproof layer will be guaranteed for other particles lying above and interlocked into these particles which effect is favourable in the consideration of increasing of inner shear resistance. Underneath there is a geogrid-crushed stone composite layer which hinders vertical and horizontal re-arrangement of particles. Geogrid creates a so-called clamped "quilt" in which there is determined and in forced manner materialized acting together of granular material particles.

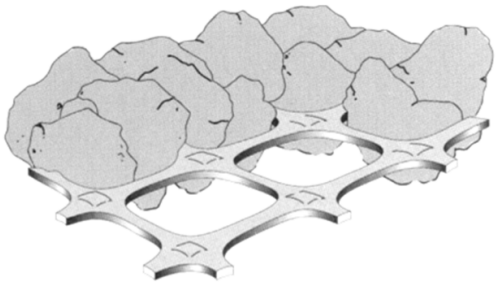


Figure 1. Crushed stone particles are wedged into the aperture of geogrid [15]

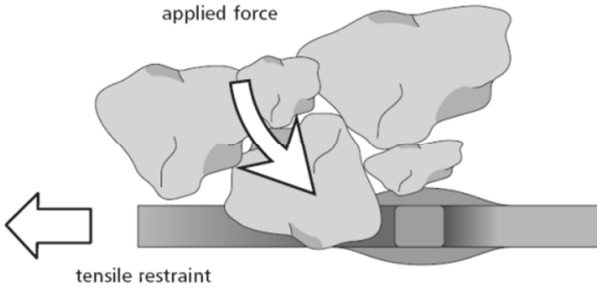


Figure 2. Interlocking effect [15]

In this way in the ballast floating, dynamic loaded railway track will geometrically be more stable and more resistant against evolving of settlement faults. Stresses arise in the ribs and junctions of the geogrid due to vehicle load, the geogrid can offer resistance against these stresses with tensile strength and low strain. Tensile strength should be adequate high, but failure strain should be acceptable low, because of the load bearing with small strain. The latter property is important because the geogrid should bear

[3] More precisely analysis of this field exceeds the frames of this paper, but it is a significantly important research topic.

[4] Naturally geogrid-reinforced railway ballast can't help abolish all the speed restriction sections, because some of them has reasons for that's abolishment this kind of stabilisation alone is inadequate.

adequate magnitude load, i.e. it can't keep out of loads way. For example non-woven geotextile can't reinforce granular material because of elongation [4].

It was a specific formulated assumption from MÁV (Hungarian Railways) as railway line operator and geogrid/geocomposite manufacturers that geogrids as well as geocomposites under railway ballast solve geometrical stabilisation of railway tracks and with this method is low-cost and adequate for simply abolishment of local track faults (e.g. water pockets). This paper deals with only the evaluation of the railway track geometry stabilisation effect of geogrid layers under ballast on the basis of laboratory tests.

2. Summary of the results of international publications

It has been mentioned that in the papers [12, 13] reduced scale assemblies were used for the laboratory tests, however the results of these measurements are queriable, it has to be highlighted that they didn't use crushed stone and geogrids/geocomposites which adequate for real railway construction. The scale of different structural elements was not the same in their laboratory tests (M 1:12 for crushed stone ballast particles, M 1:10 for geogrid thickness, M 1:3 for geogrid aperture size, M 1:36...M 1:9 for load plate size compared to a real sleeper' loading face, etc.) With these comparison and laboratory tests can't be achieved realistic behaviour of specimens, because there is unknown distortion (it would be much better than using e.g. M 1:10 scale for all elements). The most unbelievable result is that the most effective reinforcement was measured close to the loading plate (at a very small depth) [13]. This setting is impossible because of the technology; in other respects it is true that the interlocking effect has the largest value on the plane of a geogrid layer, but for this enough soil covering depth is needed. In the authors's view there were not enough soil depths in the measurements of the papers [12, 13].

The measurements and its results described in [14] showed that the largest reduction of settlements could be obtained with three geogrid/geocomposite layers (one would be between the subbase and the subgrade soil; one would be in the subbase, and one would be under the subballast). It has to be mentioned that this kind of ballast reinforcement can only be used for construction works of new railway tracks because of the very short allowed time of hold-up of the track. Local track faults can be exception[5][6]. The most effective solution for geometry stabilisation of railway track with geogrid layers under ballast is using only one layer geogrid between railway ballast and protective layer or subgrade for longer sections during ballast cleaning works or full ballast change.

On the basis of [7, 8] it is worth considering results which show that plastic deformation of ballast layer in horizontal and vertical planes too can be achieved with geogrid/geocomposite reinforcement. It is queriable that using with only geotextile layer better settlement reduction can be obtained than with only geogrid layer, because of the elongation of geotextiles.

[5] In this case it has to be considered that reason of track faults is related to sub- or superstructure because in case of track faults due to substructure defect, only using more geogrid layers can't help.

[6] The newest foreign experiences show that only one geogrid layer under railway ballast is effective, because the first compacted crushed stone layer above the bottom geogrid layer hinders the wedge of other particles into the upper laid down geogrid's aperture.

No one of the cited publications mentioned how many times they repeated the measurements for the published results, in the author's view it is an important deficiency.

Summarizing results of laboratory test of the international publications it can unequivocally be stated that geogrid/geocomposite-reinforced ballast is an effective solution for geometry stabilisation of railway track, but no one determined the property of this kind of reinforcements that how geogrid/geocomposite layers can change inner shear resistance of the original unreinforced ballast material as a function of vertical distance from geosynthetic layer. A great "hole" was left in the research of this topic, because no one can surely certify how much the increasing factor of inner shear resistance of railway ballast in case of using geogrids/geocomposites and how much the depth of the active reinforcement. For this investigation a multi-level shear box with minimum 1.0x1.0x1.0 m dimensions should be utilized which is from 10 cm high, on each other movable-slippable frames. Using this multi-level shear box inner shear resistance of ballast material with and without geogrid/geocomposite can be determined[7].

3. Laboratory tests

3.1. Aims of laboratory tests

In the author's view as well as on the basis of international publications geogrid layer under railway ballast can stabilise geometry of railway track and reinforce the load-carrying layer structure. For this there should be interlocking effect (Figure 1 and 2). Into geogrid's aperture wedged crushed stone and the upper others wedged into them increase significantly the inner shear resistance of railway ballast, the layer structure can resist better against outer loads. The acting together of crushed stone and geogrid is not totally known in the geogrid's plane. There is no information and results of laboratory test how the effect of geogrid decreases in the ballast material and where the border of effective acting together is.

The aim of the author's laboratory tests is to determine forces needed to push frames in several shearing planes in case of different parameters (e.g. type of geogrid, depth of railway ballast, sharpness of crushed stone particles, compaction level of ballast, elasticity (strength) of lower layers.

On the basis of assumptions geogrid layer under railway ballast can

- reduces slumping of the ballast bed's shoulder due to the vibrating effect of the railway load, in this way it holds better ballast resistance in cross direction,
- increases the inner shear resistance and load-carrying capacity of railway ballast.

Dimensional faults (direction, settlement, twist) will be decreased due to above effects. Because of less track faults less regulation works will be needed, in this way it has significant national economic result. If the behaviour of geogrid/geocomposite-reinforced ballast material is better known, its needed depth can be more precisely determined, it is an economic task too.

[7] For the present the authors did measurements in the plane of geogrid and in ballast material above geogrid.

3.2. Interlocking effect

Interlocking effect as a function of the distance from the geogrid layer is not known. It can probably be stated that the greater is the distance from geogrid layer the less is this effect. It is an adequate approximation that three zones are supposed (Figure 3).

The furthest zone from the geogrid layer is ZONE NC, here the effect of interlocking is the minimal. Behaviour of ballast particles is influenced by their interaction. The second zone is the transition zone (ZONE TC), here predominates the interlocking effect but the nearer to the upper plane of the ballast the smaller its value. The supposed function is non-linear (Figure 4). The third zone (ZONE FC) is located next to geogrid/geocomposite layer. Here the interlocking effect is the maximal.

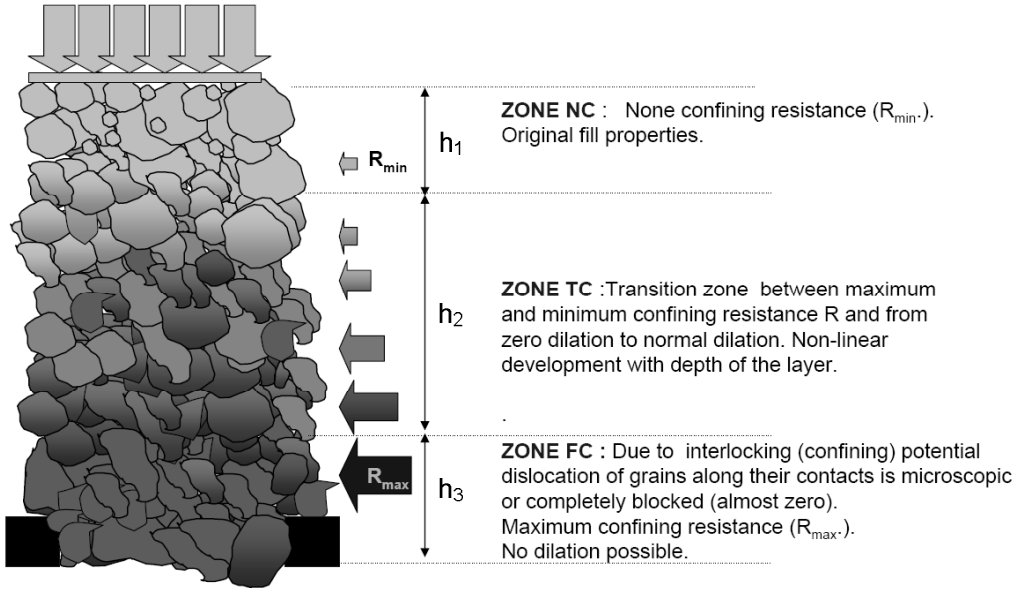


Figure 3. Hypothetical zones of confining resistance (interlocking effect) [11]

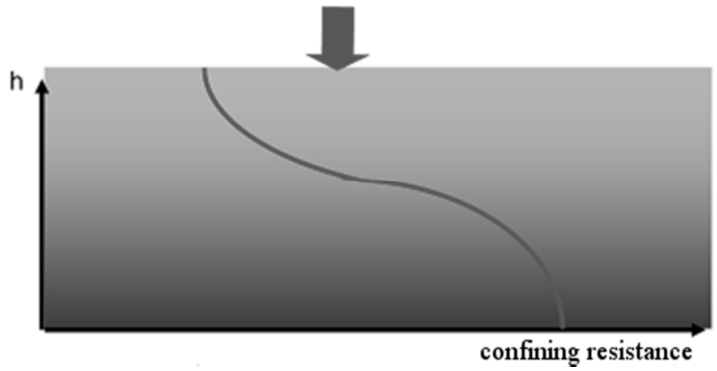


Figure 4. Supposed shape of confining resistance as a function of distance from geogrid layer [17]

3.3. Required tests determined in standards

Prescriptions of two valid standards [2, 3] have to be used, however only certain regulations from these can be utilized because of the planned special tests (multi-level shear box tests).

3.4. Method and apparatus of the laboratory tests

There should be such shear box tests to determine confining resistance as a function of distance from geogrid layer, which can give results (shear resistance) in sections of several heights of railway ballast. From this fact a special multi-level shear box had to be developed. The results are influenced by several parameters, in this way a number of measurements should be done. It is very important to ensure quasi same circumstances in the same measurement series. It means that at an identical measurement series all the parameters are the same, only the plane of shear changes (Figure 5).

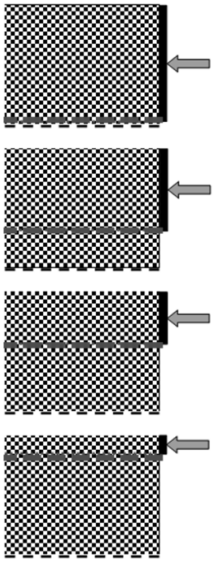


Figure 5. Shear tests in different shear planes [17]

The variables of the tests are the follows:

- elasticity (strength) of below support layer,
- type of geogrid (most important properties: aperture, elongation modulus),
- properties of ballast (grain-size distribution, shape of particles, fresh or recycled material),
- depth of ballast layer,
- compaction level of ballast material,
- loading on the top of layer structure.

The conventional two-frame shear box isn't adequate apparatus, because it always works in the same plane. For the author's task special multi-level shear box is needed, which is divided more frames vertically. The shear box consists of a lower frame and nine upper storey frames (Figure 6).

The area of shear box is 1.0x1.0 m, and its height is 1.0 m. Figure 7 shows one kind of layer structure set-up.

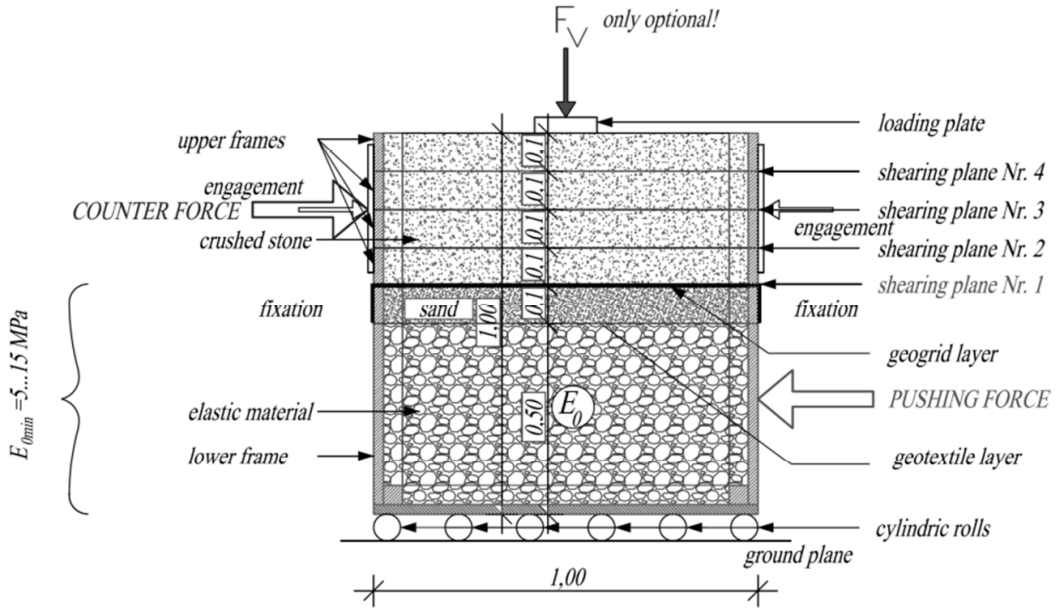


Figure 6. Principle plan of multi-level shear box, shearing in shearing plane Nr. 1 [6]

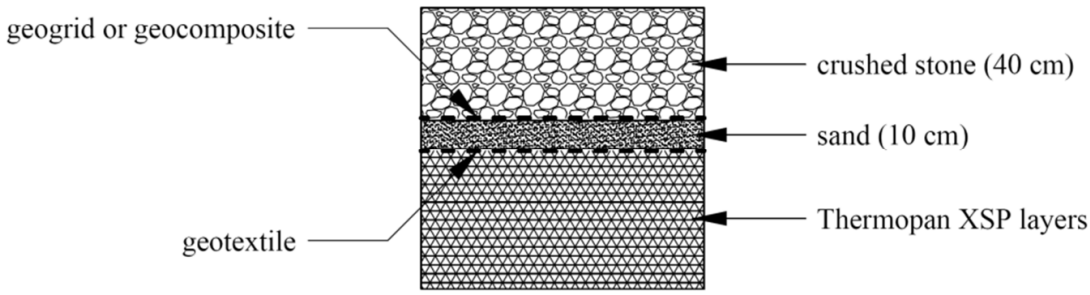


Figure 7. One kind of layer structure set-up [6]

In the lower frame there should be elastic material with low load-carrying capacity. It can be e.g. Thermopan XSP plates. The E_2 modulus of this layer should be determined with load plate test. The elasticity of the layer from Thermopan XSP plates changes if its height is changed (the concrete floor of the laboratory and steel plates used in the bottom of lower frame don't influence this value because of their approximately infinity elasticity). The very low E_2=5… 15 MPa modulus can be achieved by using 40… 50 cm height Thermopan XSP sheets.

The second layer from the bottom is sand with 10 cm height laid on geotextile layer. This layer helps the crushed stone particles penetrate into the aperture of geogrid and protects Thermopan XSP sheets against sharp edges of crushed stone. On the top of sand layer one layer of geogrid or geocomposite (geotextile+geogrid) is laid. It is the plane until the upper layer structure should be disassembled and from the new layer structure should be reconstructed. As the moving rear edge[8] of the simple geogrid layer or the geogrid in geocomposite has to be fixed to the outer wall of shear box. It symbolises that geogrid is the clamped into railway ballast in real situations. Plane of

[8] In the author's laboratory tests both edge of geogrids/geocomposites perpendicular to shearing's direction were fixed.

geogrid is the shearing plane Nr. 1. Below the lower frame there are cylindrical rolls, on which the lower frame can be pushed by horizontal force.

Upon the geogrid/geocomposite crushed stone ballast of 40 cm height is laid down. Crushed stone ballast is bordered by four frames of 10 cm height. Over the plane of geogrid/geocomposite layer every plane of frame meet is shearing plane. Over and below of the chosen shearing plane frames should be engaged together because of their simultaneous moving. For example if the shearing plane Nr. 2 is chosen the lower seven frames are engaged together as well as the upper three frames in the same way, etc.

It is very important to ensure quasi same circumstances in the same measurement series. It means that in every test series the compaction level of ballast is the same, but there isn't such apparatus with which measures the density of ballast material in the shear box. In this way below things have to be done:

- always the same compaction apparatus should be used,
- always the same number of compaction passes should be utilized.

In each shear test parameters should be known or measured as follows:

- E_2 modulus of below support layer,
- grain-size distribution and shape of ballast particles,
- depth of ballast,
- value of static loading on the upper plane of ballast (if there is),
- value of horizontal force need to move the frames below the actual shearing plane (pushing force),
- value of horizontal force needed to strut the frames above the actual shearing plane (counter force),
- movement of the frames below the actual shearing plane,
- value of vertical force needed to counterforce the frame lift due to ballast dilation[9].

Figure 8 shows the planned and manufactured shear box.

Figure 8. For the tests prepared multi-level shear box [17]

[9] Forces were recorded but the set-up of shear box doesn't influence the inner shear resistance of ballast, because this vertical force stays frame on the original shearing planes, in this way it was neglected.

On both sides (which are parallel to shearing direction) of upper five frames there are windows made of plexi with 200x60 mm dimensions. Through these windows the movement of crushed stone particles and sand can be monitored during shearing tests and it can be determined whether particle movement and rotation influences particle layers in below shearing planes. Setting frames on each other steel L-profiles ensure. Frames laid on each other can be fixed by M12 screws with vertical axes. Those two frames between them there is shearing plane aren't naturally clamped together.

The authors considered:

- compaction level with two values (uncompacted, compacted),
- two types of geogrid and two types of geocomposite (geogrid+geotextile) (namely geogrid type 1 and 2, as well as geocomposite type 1 and 2),
- ballast layer with E_2 modulus of 7.2 MPa,
- fresh ballast material (with sharp edges),
- constant ballast depth of 40 cm,
- zero vertical static loading,
- four shearing planes

during up to now completed laboratory tests.

3.5. Laboratory tests and their results

Respect to above written the following tests were done:

- grain-size distribution test and shape test of ballast particles,
- measure bedding property of support layer structure,
- measure friction resistance between empty box frames during shearing,
- inner shear resistance without vertical static loading,
    - uncompacted ballast , without geogrid,
    - uncompacted ballast, with geocomposite type 1,
    - compacted ballast without geogrid,
    - compacted ballast with four types of geogrid/geocomposite.

Three measurements were done in case of each assembly for the four shearing planes to characterize the inner shear resistance.

Crushed stone ballast material was given by mine of KŐKA Kő- és Kavicsbányászati Kft. from Komló (Hungary). Ballast material complies with the requirements of standard [1]. Because of limited space detailed tests can't be published in this paper.

3.5.1. Bedding property of support layer structure

Bedding of support layer structure was obtained by Thermopan XSP sheets of 50 cm height. Strength (E_2 modulus) of this structure was measured by static load plate test. Two measurements were done, and in the second loading cycle average settlement s_2=9.4 mm, therefore E_2=67.5/s_2=7.2 MPa. In each assembly E_2=7.2 MPa modulus was used.

3.5.2. Friction resistance between empty box frames during shearing

There is friction resistance in shearing planes of box between frames. Their values were determined by measurements with two repeats:

- shearing plane Nr. 4: 0.265 kN,
- shearing plane Nr. 3: 0.462 kN,
- shearing plane Nr. 2: 0.664 kN,
- shearing plane Nr. 1: 0.865 kN.

3.5.3. Investigations of different layer structure

During all laboratory tests vertical static loading was zero, E_2 modulus of below support layer was 7.2 MPa, depth of ballast layer was 40 cm, depth of sand layer below the ballast was 10 cm, the geotextile between sand layer and Thermopan sheets was Secutex 151 GRK 3.

3.5.3.1. Properties of geogrids/geocomposites used in laboratory tests

Four types of geosynthetics were used in the laboratory tests:

- geogrid type 1,
- geocomposite type 1 (geogrid type 1 + geotextile),
- geogrid type 2,
- geocomposite type 2 (geogrid type 2 + geotextile).

Geogrid type 1 and geogrid in geocomposite type 1 are extruded, geogrid type 2 and geogrid in geocomposite type 2 are welded. Geometric properties of geogrids/geocomposites can be seen in Figure 9 and in Table 1, mechanic properties of them are consisted Table 2 and 3.

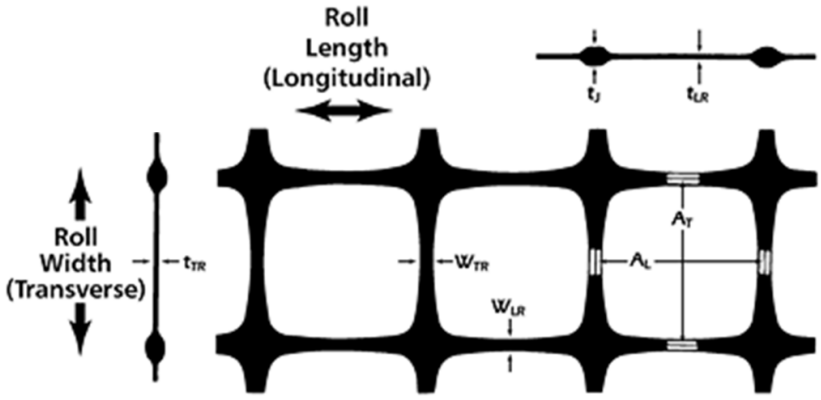


Figure 9. Definition of geometric properties of geogrids [17]

Table 1. Geometric properties of geogrids [6]

Geogrid/geo-composite type	**A_L (mm)**	**A_T (mm)**	**W_{LR} (mm)**	**W_{TR} (mm)**	**t_J (mm)**	**t_{LR} (mm)**	**t_{TR} (mm)**
Geogrid type 1	65.0	65.0	4.0	4.0	7.0	1.7	1.5
Geocomposite type 1	65.0	65.0	4.0	4.0	7.0	1.7	1.5
Geogrid type 2	80.0	80.0	8.8	8.2	2.1	1.4	1.4
Geocomposite type 2	80.0	80.0	8.8	8.2	2.1	1.4	1.4

Table 2. Mechanical properties of geogrids 1. [6]

Geogrid type	**Material**	**Uniaxial/ Biaxial**	**Ultimate strength**		**Strength at 2 % elongation**		**Ultimate elongation**	
			MD (kN/m)	**XMD (kN/m)**	**MD (kN/m)**	**XMD (kN/m)**	**MD (kN/m)**	**XMD (kN/m)**
Geogrid type 1	PP	Biaxial	30	30	11	12	N.A.	N.A.
Geocompo-site type 1	PP	Biaxial	30	30	11	12	N.A.	N.A.
Geogrid type 2	PP	Biaxial	30	30	12	12	N.A.	N.A.
Geocompo-site type 2	PP	Biaxial	30	30	12	12	N.A.	N.A.

Table 3. Mechanical properties of geogrids 2. [6]

Geotextiles and geotextiles in geocomposites	**Puncture resistance (N)**	**Ultimate strength**		**Ultimate elongation**		**Permeability (m/s)**	**Permeability (l/sm^2)**	**Unit weight (kg/m^2)**	**Effective opening size (mm)**
		MD (kN/m)	**XMD (kN/m)**	**MD (%)**	**XMD (%)**				
Geocompo-site type 1	>1500	N.A.	N.A.	N.A.	N.A.	0.135	135	0.16	0.125
Geocompo-site type 2	1670	6	11	60	40	0.11	110	0.15	0.13
Naue Secutex 150 GRK 3	1670	6	11	50	30	0.09	90	0.15	0.08

3.5.3.2. Execution of laboratory tests

3.5.3.2.1. Uncompacted layer structure without geogrid

In this series ballast was uncompacted. Shearing test was started in shearing plane Nr. 4, and then followed Nr. 3, 2 and 1. In each shearing frame moving was approximately 30…80 mm. This was generally enough to occur stationary pushing force, which shouldn't be increased to slip frames permanently. This frame moving of 30… 80 mm didn't change position of crushed stone particles in the lower layers[10].

Pushing force was increased with speed of 20 kN/min. In each shearing plane three measurements were done. After each series ballast material had been removed from the box, and then it was reconstructed[11].

3.5.3.2.2. Compacted layer structure without geogrid

Tests were done introduced in Section 3.5.3.2.1. Ballast material was compacted in 20 cm height layers by an L-2/C vibrator (68 kg, 1.1 kW power, 3000 1/min nominal vibration frequency, 500x500 mm vibration plane). Achieving same density (compaction level) always the same number of compaction passes should be utilized (on two lanes with three passes).

Figures 10-13 illustrate results of measurements related to compacted layer structure without geogrid.

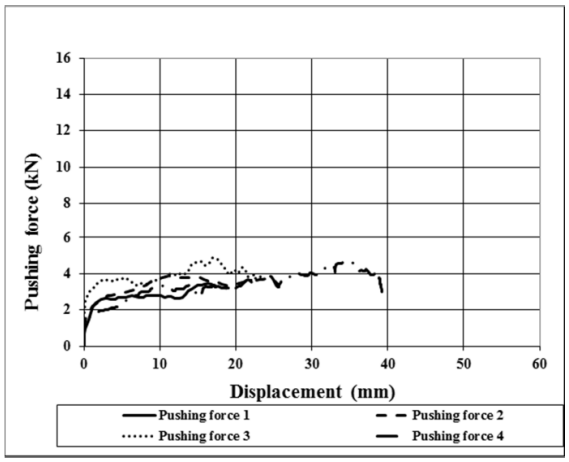


Figure 10. Pushing force-displacement diagram, compacted layer structure without geogrid, shearing plane Nr. 4. [17]

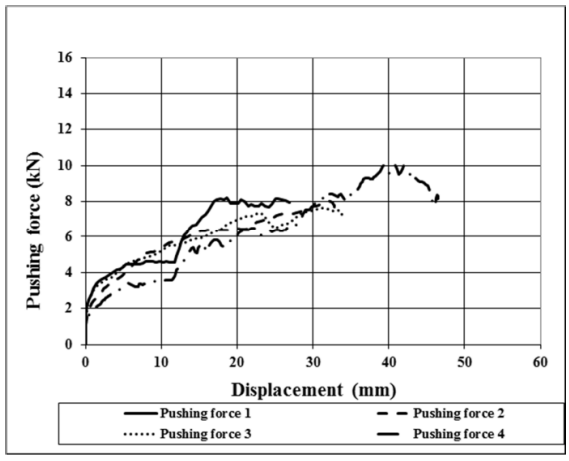


Figure 11. Pushing force-displacement diagram, compacted layer structure without geogrid, shearing plane Nr. 3. [17]

[10] It was monitored through plexi windows.

[11] If geogrid/geocomposite-reinforced layer structure was investigated, after each series new geogrid/geocomposite was built-in.

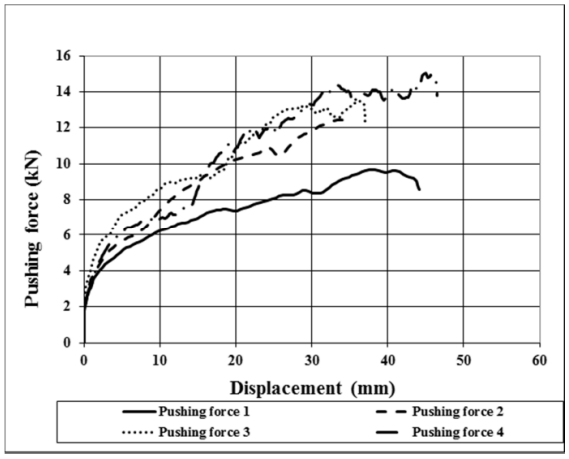


Figure 12. Pushing force-displacement diagram, compacted layer structure without geogrid, shearing plane Nr. 2. [17]

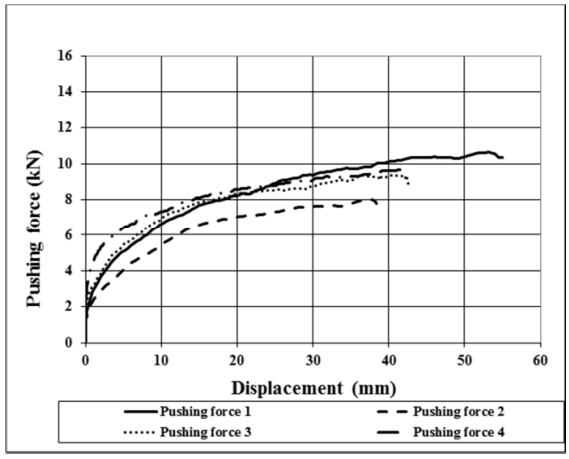


Figure 13. Pushing force-displacement diagram, compacted layer structure without geogrid, shearing plane Nr. 1. [17]

3.5.3.2.3. Uncompacted and compacted layer structures with geogrid/geocomposite

Geogrids and geocomposites were set in shear box as written in Section 3.4. In this paper only diagram related to geocomposite type 1 are shown in Figure 14-17.

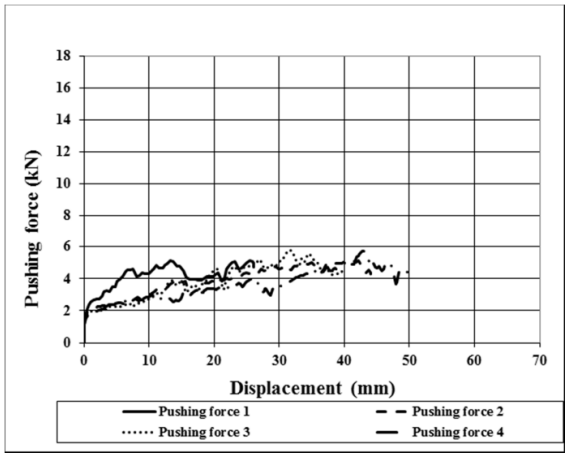


Figure 14. Pushing force-displacement diagram, compacted layer structure with geocomposite type 1, shearing plane Nr. 4. [17]

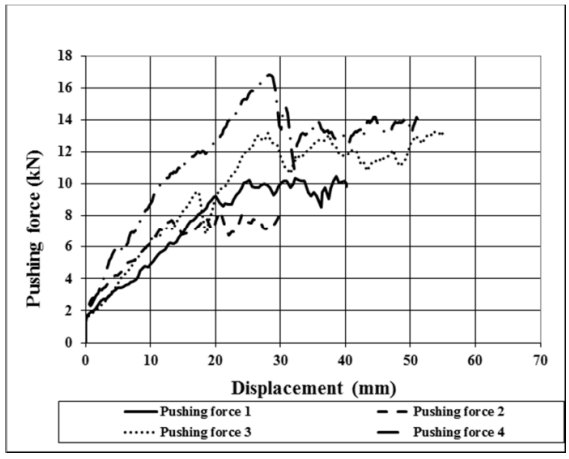


Figure 15. Pushing force-displacement diagram, compacted layer structure with geocomposite type 1, shearing plane Nr. 3. [17]

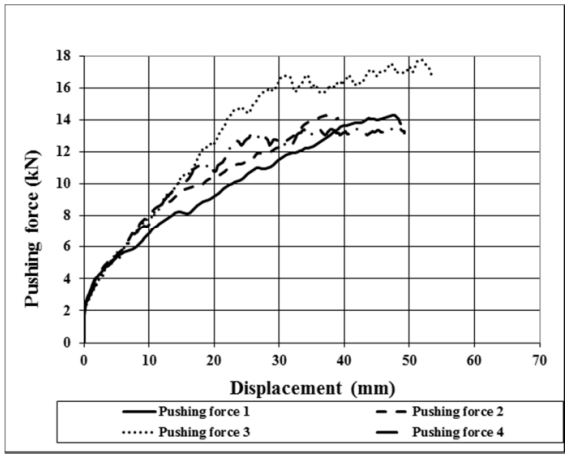


Figure 16. Pushing force-displacement diagram, compacted layer structure with geocomposite type 1, shearing plane Nr. 2. [17]

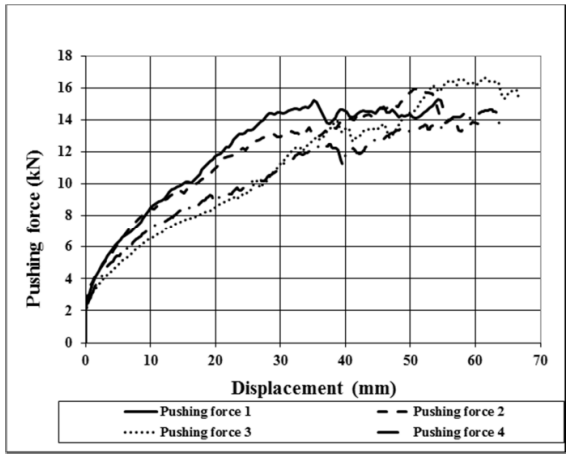


Figure 17. Pushing force-displacement diagram, compacted layer structure with geocomposite type 1, shearing plane Nr. 1. [17]

Because of limited space other diagrams can't be published in this paper. Results of all laboratory tests are consisted in Section 3.5.3.3.2.

3.5.3.3. Methods of measured data's process and results of laboratory tests

3.5.3.3.1. Methods of measured data's process

In a pushing force-displacement diagram it can be seen that they show bounded above graphs. This boundedness characterizes the particle-mechanical behaviour that relates to inner shear resistance of railway ballast. It is obvious that aggregates like railway ballast (crushed stone with sharp edged particle) can't represent exact inner shear resistance neither in laboratory. Measured data have some deviation because of random distribution of irregular shaped particles. Pushing force-displacement diagram related to shearing in a certain shearing plane can be evaluated by mathematical statistic methods which gives inner shear resistance of ballast in the certain height. These values can be calculated by determination of average of stationary pushing forces[12] related to different cases (uncompacted, compacted, with or without geogrid, etc.).

In the case of pushing force and related counter force (Figure 6) differed than 15 %, measurements should be repeated, therefore these values aren't contained in Section 3.5.3.3.2.

Recorded pushing forces weren't corrected by friction resistance shown in Section 3.5.2. This neglection can be used because friction resistance values are very low correlated to measured inner shear resistance values.

3.5.3.3.2. Result of laboratory tests

Inner shear resistance functions of crushed stone railway ballast were determined by using multi-level shear box tests in case of uncompacted and compacted

[12] It means interval in which additional pushing force shouldn't be to increase displacement.

(Section 3.5.3.2.2.) aggregates without or with geogrid/geocomposite reinforcement. Table 4 summarizes all measured data.

Table 4. Measured pushing forces [kN] related to inner shear resistance values of railway ballast as a function of distance from geogrid/geocomposite layer [6]

Distance from geogrid/geocomposite layer (cm)	No. of measurement	Uncompacted		Compacted				
		Without geogrid	With geocomposite type 1	Without geogrid	With geocomposite type 1	With geogrid type 1	With geogrid type 2	With geocomposite type 2
0	**1**	7.37	16.39	10.38	14.54	16.74	15.27	15.28
	2	8.52	7.45	7.64	14.16	17.93	14.36	14.32
	3	-	11.31	9.03	16.11	18.49	15.86	13.50
	4	-	10.36	9.38	13.86	-	-	-
	Min.	7.37	7.45	7.64	13.86	16.74	14.36	13.50
	Max.	8.52	16.39	10.38	16.11	18.49	15.86	15.28
	Avg.	**7.95**	**11.38**	**9.11**	**14.67**	**17.72**	**15.16**	**14.36**
	Dev.	0.81	3.72	1.13	1.00	0.89	0.76	0.89
10	**1**	6.72	6.59	9.61	14.08	16.02	18.43	17.90
	2	6.05	6.40	12.27	14.12	15.51	16.19	16.20
	3	-	8.99	12.76	16.67	15.62	16.64	16.43
	4	-	9.54	14.02	13.23	-	-	-
	Min.	6.05	6.40	9.61	13.23	15.51	16.19	16.20
	Max.	6.72	9.54	14.02	16.67	16.02	18.43	17.90
	Avg.	**6.39**	**7.88**	**12.17**	**14.53**	**15.72**	**17.09**	**16.84**
	Dev.	0.47	1.62	1.86	1.49	0.27	1.19	0.92
20	**1**	3.28	3.32	7.92	9.84	9.91	13.89	10.63
	2	3.79	3.52	7.79	7.57	9.74	10.09	11.73
	3	-	6.18	7.49	12.16	11.79	16.86	12.07
	4	-	6.83	9.74	13.19	-	-	-
	Min.	3.28	3.32	7.49	7.57	9.74	10.09	10.63
	Max.	3.79	6.83	9.74	13.19	11.79	16.86	12.07
	Avg.	**3.54**	**4.96**	**8.24**	**10.69**	**10.48**	**13.61**	**11.47**
	Dev.	0.36	1.80	1.02	2.51	1.14	3.39	0.75
30	**1**	1.67	2.64	3.28	4.86	3.35	4.93	4.15
	2	1.76	1.83	3.62	4.82	5.53	4.31	4.94
	3	-	3.52	4.27	5.15	3.15	4.16	6.02
	4	-	2.58	3.93	4.81	-	-	-
	Min.	1.67	1.83	3.28	4.81	3.15	4.16	4.15
	Max.	1.76	3.52	4.27	5.15	5.53	4.93	6.02
	Avg.	**1.72**	**2.64**	**3.78**	**4.91**	**4.01**	**4.47**	**5.04**
	Dev.	0.06	0.69	0.42	0.16	1.32	0.41	0.94

Averages of pushing forces are shown in Figure 18. Polynomial regression functions were utilized to evaluate the results, their equations and R^2 values are given in Table 5 and 6. Boundary condition was considered that on the upper surface of ballast (40 cm distance from geogrid layer) shearing can't be interpreted.

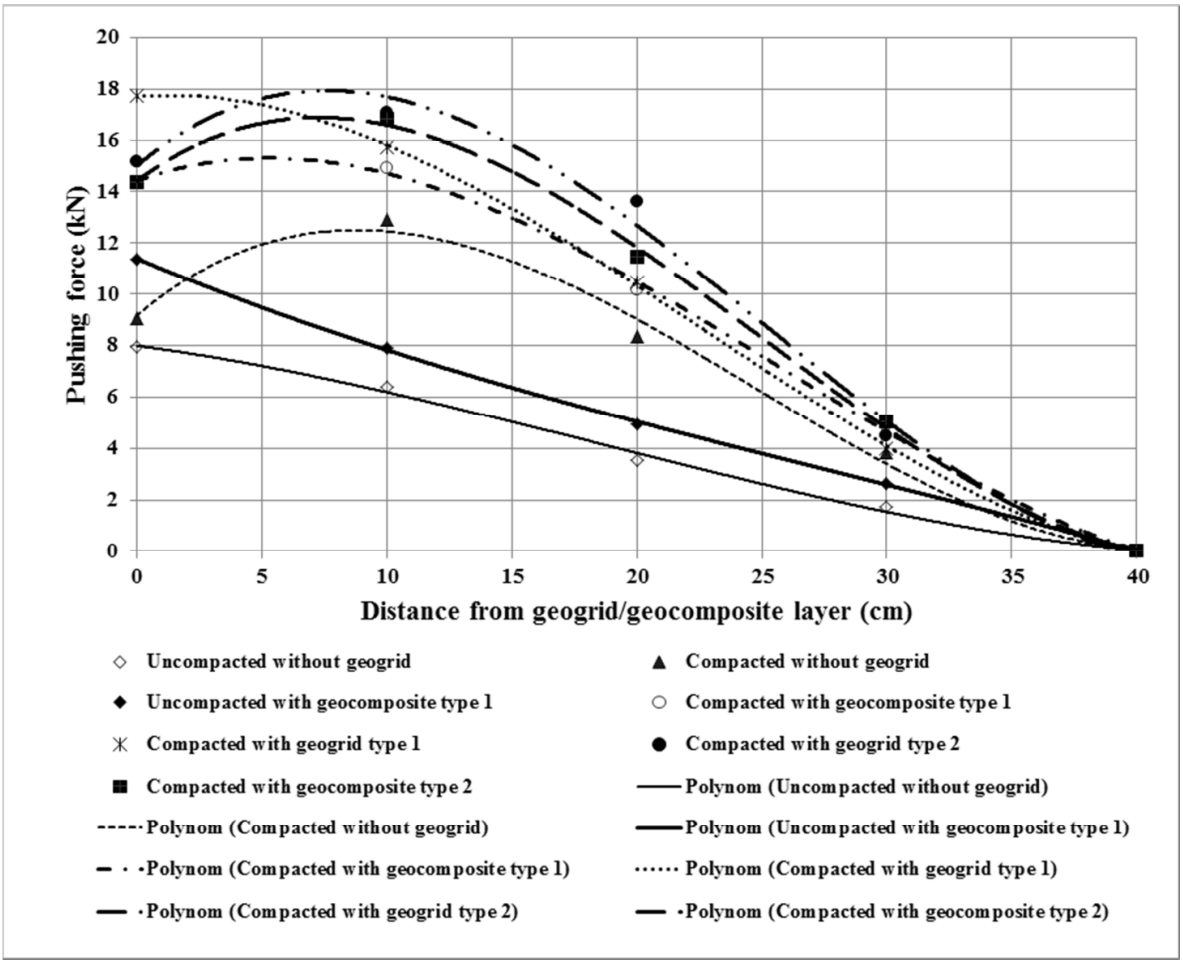


Figure 18. Averages of measured pushing forces as a function of distance from geogrid/geocomposite layer [6]

Table 5. Factors and R^2 values of cubic polynomial regression functions 1 [6]

	Uncompacted without geogrid	**Compacted without geogrid**	**Uncompacted with geocomposite type 1**	**Compacted with geocomposite type 1**
const.	7.991357	9.153036	11.390893	14.426460
x	–0.126196	0.820080	–0.410533	0.349165
x^2	–0.006461	–0.056529	0.006161	–0.036846
x^3	0.000116	0.000759	–0.000075	0.000478
R^2	0.9965	0.9915	0.9998	0.9985

Table 6. Factors and R^2 values of cubic polynomial regression functions 2 [6]

	Compacted with geogrid type 1	**Compacted with geogrid type 2**	**Compacted with geocomposite type 2**
const.	17.693390	15.008614	14.426210
x	0.086873	0.820204	0.716988
x^2	–0.032234	–0.063581	–0.057716
x^3	0.000475	0.000840	0.000770
R^2	0.9998	0.9927	0.9986

Conclusions can be derived from results are written in Section 3.6.

After removing layer structure from shear box significant failures can be seen on geogrid type 1 and geogrid type 2. Geogrids with welded junctions (geogrid type 2) don't ensure adequate solution because of their vulnerability[13]. In case of extruded geogrid (geogrid type 2) only smaller cracks can be observed.

Rotational resistance in geogrid plane was investigated in case of geogrid type 1 and geogrid type 2. They are characterized by rotations (degrees) due to moments of 1.0 Nm[14]:

$$\varphi_{geogrid\_type\_1.,1\_Nm} = 1{,}38°, \tag{1}$$

$$\varphi_{geogrid\_type\_2.,1\_Nm} = 19{,}02°, \tag{2}$$

Increasing factors as a function of distance from geogrid layer were defined by using polynomial regression functions of pushing forces (inner shear resistance), their mechanical meaning are the following:

- increasing factor "A": inner shear resistance related to geogrid/geocomposite-reinforced compacted ballast divided by inner shear resistance related to unreinforced compacted ballast (effect of geogrid/geocomposite reinforcement in compacted ballast),
- increasing factor "B": inner shear resistance related to geogrid/geocomposite-reinforced compacted ballast divided by inner shear resistance related to geogrid/geocomposite-reinforced uncompacted ballast (effect of compaction using geogrid/geocomposite reinforcement),
- increasing factor "C": inner shear resistance related to geocomposite-reinforced compacted ballast divided by inner shear resistance related to geogrid-reinforced uncompacted ballast (effect of geotextile using geogrid/geocomposite reinforcement in compacted ballast),
- increasing factor "D inner shear resistance related to geogrid/geocomposite-reinforced uncompacted ballast divided by inner shear resistance related to

[13] Geogrid type 2 and geocomposite type 2 were investigated only as a probe because they are not used for geometry stabilisation of ballasted railway track.

[14] These results are from FEM modelling with AxisVM 11 software. This kind of tests was not done by authors.

unreinforced uncompacted ballast (effect of geogrid/geocomposite reinforcement in uncompacted ballast),

- increasing factor "E": inner shear resistance related to unreinforced compacted ballast divided by inner shear resistance related to unreinforced uncompacted ballast (effect of compaction in unreinforced ballast).

Increasing factors "A" to "E" as a function of distance (0…30 cm[15]) from geogrid/geocomposite layer are shown in Figure 19-23. The ordinates of graphs were calculated by ratio of specific equations. Because of limited space factors and R^2 values of regression functions can't be published in this paper, but it should be noticed that all R^2 values are higher than 0.97.

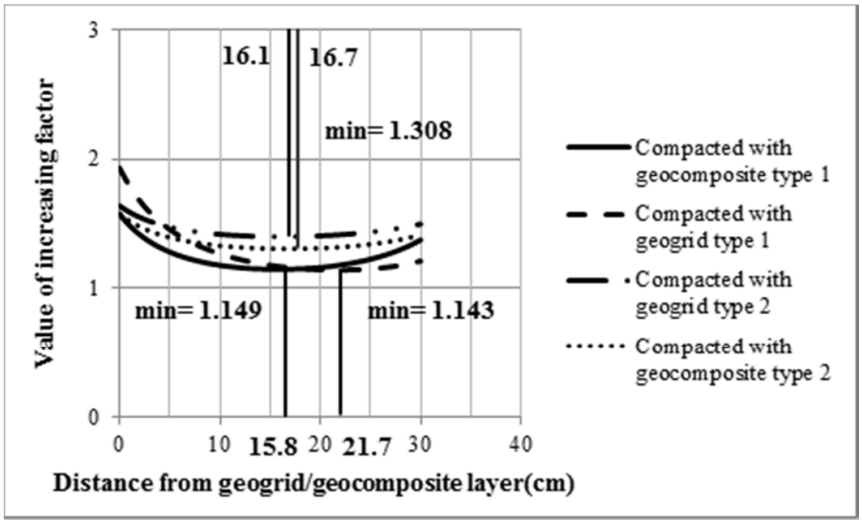


Figure 19. Value of increasing factor "A" [6]

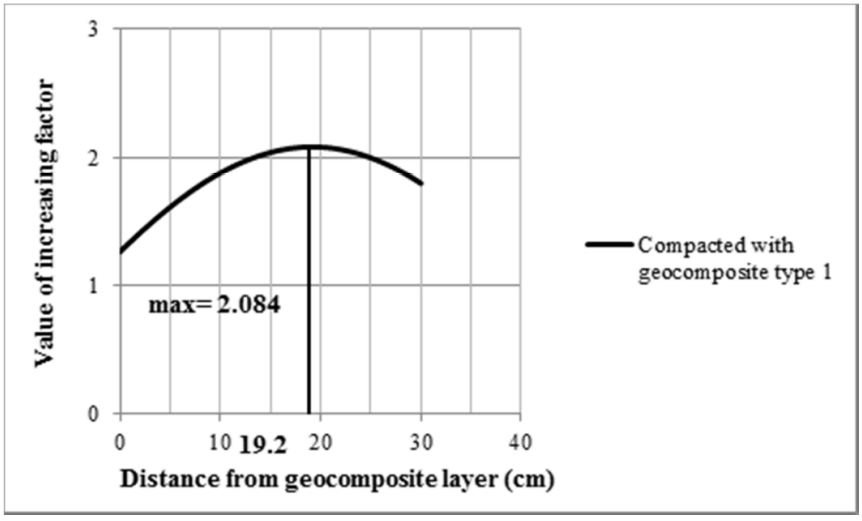


Figure 20. Value of increasing factor "B" [6]

[15] In planes which are 30…40 cm from geogrid/geocomposite layer there are no measured data. Referred to Section 3.5.3.3.2. all increasing factors can be set to 1.0 in plane 40 cm from geogrid/geocomposite layer.

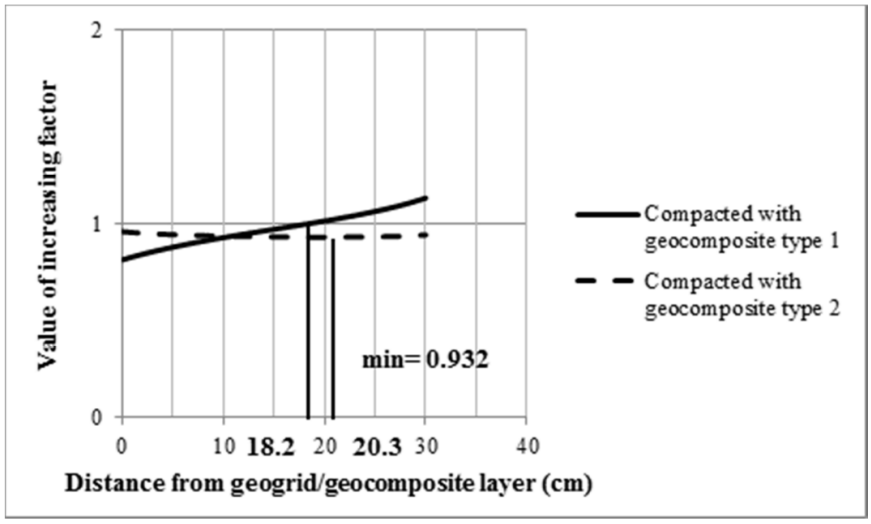


Figure 21. Value of increasing factor "C" [6]

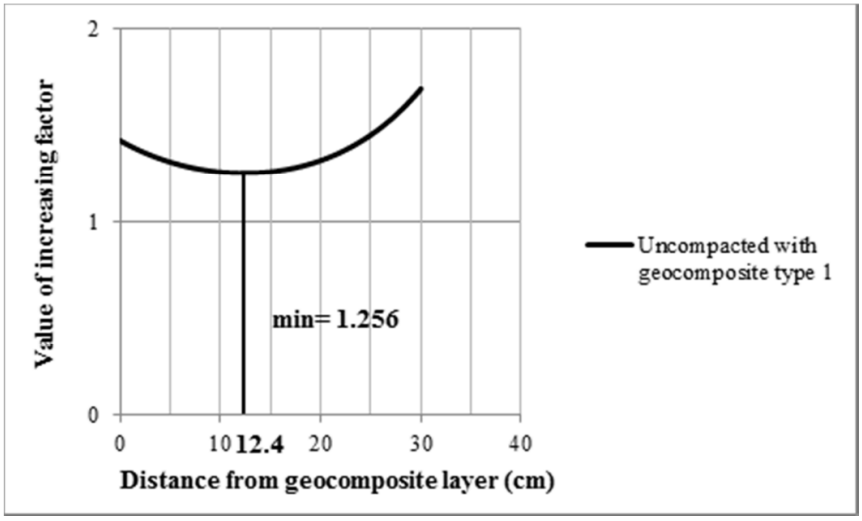


Figure 22. Value of increasing factor "D" [6]

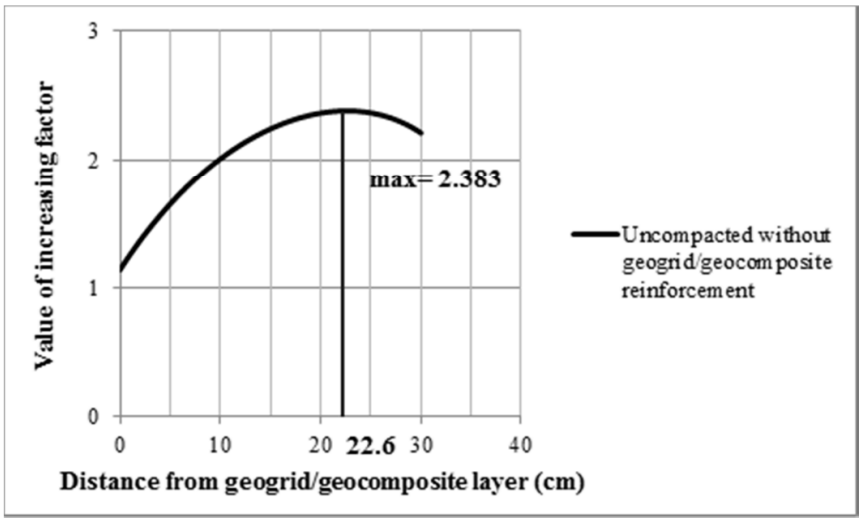


Figure 23. Value of increasing factor "E" [6]

3.6. Conclusions derived from results of laboratory tests

It can unequivocally be stated on the basis of results written in Section 3.5.3.3.2 that multi-level shear box is adequate for determining inner shear resistance of granular aggregates, e.g. crushed stone railway ballast. Using this data and considering boundary conditions regression functions of inner shear resistance can be determined as a function of distance from geogrid/geocomposite layer. It should be noticed that values of these functions are approximate but reliable only in the height of shearing planes.

In the consideration of measured data of multi-level shear box tests it can also unequivocally be stated that adequate type of geogrid/geocomposite under ballast can increase inner shear resistance of railway ballast aggregate in the following ways:

- maximum value of inner shear resistance function is not in the plane of geogrid/geocomposite (except geogrid type 1), but 0...10 cm above this plane,
- in case of geocomposite type 1 and geocomposite type 2 inner shear resistance is decreased (correlated only to same manufactured product) by glued geotextile layer because it significantly hinders ballast particle's wedging into aperture of geogrids (in case of geocomposite type 1 in the 0...18 cm zone, in case of geocomposite type 2 in the whole 0...40 cm zone),
- the reason of smaller measured pushing forces in case of geogrid type 2 was the notable observed failure because of its weakness.

Increasing factors were determined, which show the following results:

- effect of geogrid/geocomposite reinforcement in compacted ballast is minimal in the 15...22 cm zone. Maximal increasing is observed in case of geogrid type 1 (1.904). In the plane of geogrid/geocomposite the most effective are geogrids without geotextile (geogrid type 2 has this advantage in the whole 0...40 cm zone correlated to its geocomposite pair). In the zone of 10...30 cm increasing factor "A" can be considered as constant.
- Compaction increases inner shear resistance of compacted layer structure constructed by geocomposite type 1 in the whole 0...30 cm zone. Maximal increasing is observed in the height of 19.2 cm (2.084). In any other heights smaller increasing can be obtained.
- Evaluate the effect of geotextile using geogrid/geocomposite reinforcement in compacted ballast increasing can only be received by using geocomposite type 1 in the zone of 18.2...30 cm. In case of geocomposite type 2 can't be achieved increasing correlated to geogrid type 2.
- In uncompacted ballast the investigated geocomposite type reinforces railway ballast aggregate in the whole 0...30 cm zone. Maximal increasing can be observed in the maximum distance from geocomposite layer. Minimal increasing is in the height of 12.4 cm (1.255), in any other heights greater increasing can be obtained.
- Compaction significantly increases inner shear resistance of uncompacted layer structure. Its maximum value is 2.380 in the height of 22.6 cm. In any other heights smaller increasing can be obtained.

It can be stated that optimal depth of compacted railway ballast is 23 cm below from the lower face of sleeper. In case of geogrid/geocomposite-reinforced compacted ballast

optimal depth would be 0…15 cm, but in the heights more far from 15 cm reinforce effect can be also obtained. In the consideration of separation of ballast and protective layer or subgrade, as well as drainage, geocomposite layer should be used – from the investigated types geocomposite type 1 is recommended –, but in this case reinforcement using geocomposite is smaller than using geogrid in the height of 0…18 cm zone. It should be noted that ballast cleaning and tamping works needed minimum 22…33 cm ballast depth between the lower face of sleeper and geogrid/geocomposite reinforcement, therefore these minimal values should be considered at design phase.

Turning-moment of unit in the plane of geogrids causes 13.78 times greater rotation of junction in case of geogrid type 1 and geocomposite type 1 than in case of geogrid type 2 and geocomposite type 2. Modulus of elasticity of geogrid type 1's material is also approximately 15 times greater than geogrid type 2's. In the consideration of these facts geogrid type 2 and geocomposite type 2 are unadequate for geometry stabilisation of ballasted railway track as reinforcement layer under ballast bed.

4. Summary and future research possibilities

This article investigated the railway track geometry stabilisation effect of geogrid layers under ballast with a specific laboratory multi-level shear box. During the laboratory tests different types of geogrid and geocomposite layers were analysed when railway ballast was uncompacted and compacted. Two types from these have not utilised for railway track geometry stabilisation yet. Inner shear resistance of railway ballast was determined in case of unreinforced and geogrid-reinforced assemblies, as well as five multiplication factors were defined which are adequate for determining inner shear resistance of reinforced and unreinforced railway ballast in consideration of different parameters.

Taking into account other parameters can help evaluation of railway track geometry stabilisation effect more precisely:

- using not only fresh, but recycled crushed stone railway ballast,
- using not only dry, but wet and oily crushed stone railway ballast,
- using different elasticity support layer,
- using different ballast depths,
- using other different geogrids/geocomposites,
- considering vertical static load on the upper surface of ballast,
- considering dynamic loads.

Acknowledgements

The authors would like to thank for help of Dénes Szekeres and Zsolt Csonka as well as the manufacturers of geogrids/geocomposites because they gave free run of geomaterials investigated in laboratory.

References

[1] MSZ EN 13450:2003: *"Kőanyaghalmazok vasúti ágyazathoz"* in Hungarian, CEN, 2003

[2] ISO 12957-1:2005: *"Geoszintetikák. A súrlódási jellemzők meghatározása. 1. rész: Közvetlen nyíróvizsgálat"* in Hungarian, CEN, 2005

[3] MSZ EN ISO 13738:2005: *"Geotextíliák és rokon termékeik. A talajból való kihúzással szembeni ellenállás meghatározása"* in Hungarian, CEN, 2005

[4] FISCHER, SZ.: *"Georácsos vasúti felépítménystabilizáció hatékonysága"* in Hungarian, XV. Nemzetközi Építéstudományi Konferencia (EMT), Csíksomlyó, Románia, 2011. június 2-5., pp. 137-144

[5] FISCHER, SZ: *"Lassújel miatti többletköltségek, és a megszüntetés költségeinek összehasonlítása"* in Hungarian, Sínek Világa, Vol. LIII (2011), No. 5, pp. 21-29

[6] FISCHER, SZ.: *"A vasúti zúzottkő ágyazat alá beépített georácsok vágánygeometriát stabilizáló hatásának vizsgálata"* in Hungarian, Ph.D. thesis, Széchenyi István Egyetem Műszaki Tudományi Kar, 2012, 148 p.

[7] INDRARATNA, B., SHAHIN, M., RUIJIKIATKAMJORN, C., CHRISTIE, D.: *Stabilisation of ballasted rail tracks and underlying soft formation soils with geosynthetic grids and drains*, ASCE Special Geotechnical Publication No. 152, Proceedings of Geo-Shanghai 2006, Shanghai, China, 2-4 June 2006, pp. 143-152

[8] INDRARATNA, B., SHAHIN, M. A., SALIM, W.: *Stabilisation of granular media and formation soil using geosynthetics* with special reference to railway engineering, Journal of Ground Improvement, Vol. 11 (2007), No. 1, pp. 27-44

[9] www.kti.hu – *"Grafikus adatbázisok, Trendek, Dugófigyelő"* in Hungarian – Közlekedéstudományi Intézet Nonprofit Kft. - http://www.kti.hu/index.php/szolgaltatasok/trendek-grafikus-adatbazis/trendek---grafikus-adatbazis [read: 12.09.2011]

[10] *"D.54 sz. Építési és pályafenntartási műszaki adatok, előírások"* in Hungarian, Part I., KÖZDOK, 1987, Budapest, 325 p.

[11] RAKOWSKI, Z., KAWALEC, J.: *Mechanically stabilized layers in roadconstruction*, not published printed conference presentation, XXVII. International Baltic Road Conference, August 2009, Riga (Latvia)

[12] RAYMOND, G. P.: *Reinforced ballast bahaviour subjected to repeated load*, Geotextiles and Geomembranes, Vol. 20 (2002), pp. 39-61

[13] RAYMOND, G., ISMAIL, I.: *The effect of geogrid reinforcement on unbound aggregates*, Geotextiles and Geomembranes, Vol. 21 (2003), pp. 355-380

[14] SHIN, E. C., KIM, D. H., DAS, B. M.: *Geogrid-reinforced railroad bed settlement due to cyclic load*, Geotechnical and Geological Engineering, Vol. 20 (2002), pp. 261-271

[15] TENSAR INTERNATIONAL LTD.: *Railways. Mechanical Stabilisation Track Ballast and Sub-ballast*, marketing issue, 2010, Blackburn, 11 p.

[16] TÖMPE, I.: *"A pályavasúti üzletág jelene és jövője"* in Hungarian, not published printed conference presentation, Új technológiák és anyagok a pályaépítésben és fenntartásban szakmai továbbképzés", 31th August – 2nd September 2011, Békéscsaba (Hungary)

[17] UNIVERSITAS-GYŐR NONPROFIT KFT.: *"Georácsok alkalmazása a vasúti zúzottkőágyazat stabilizálására"* in Hungarian, research report (written by F. Horvát, Sz. Fischer), Győr, 30th November 2010, 139 p.

[18] STAHL, M.: Interaktion Geogitter-Boden: *"Numerische Simulation und experimentelle Analyse"* in German, Ph.D. thesis, Technischen Universität Bergakademie Freiberg, 2011, 166 p.

[19] VASZARY, P.: *"A pályaromlás elmélete"*, in I. MEZEI és Id. F. HORVÁTH (eds.) Vasútépítés és pályafenntartás Part II., Magyar Államvasutak Rt., Budapest, 1999, pp. 157-160

[20] WEINREICH, Z.: *"Nagysebességű vasutak pályafenntartási kitűzése"* in Hungarian, Sínek Világa, Vol. LIII (2011), No. 6, pp. 27-31

Measurement of the diameter of the imprint based on image processing using MathCAD and the evaluation software of an industrial CT

I. Kozma[1], E. Halbritter[2]

[1,2] Department of Materials Science and Technology, Széchenyi István University,
9026 Győr, Egyetem tér 1.
kozma@sze.hu, halbritt@sze.hu

Abstract: Instead of the real photograph of a certain object its geometrical model was used to determine an unknown parameter (diameter) starting from a known parameter, using image processing. The test example allowed us to vary certain parameters of the digital image to analyse the factors influencing the precision of size determination. Basics of image processing are presented by the use of the Mathcad software. Using the Mathcad software, by a new application of well-known knowledge a method and a program were developed for the approximate determination of the center and radius (diameter) of a circle. Using the processing software of an industrial CT we demonstrated the effect of pixel size and hue level of pictures of various resolutions on the precision of the determined parameter.

Keywords: image processing, pixel number, diameter of the regression circle, Monte-Carlo method

1. Introduction

Modeling has a major rule both in the scientific area and the applied research activities in the industrial area [1, 2]. Prefabricated part for the axle stud is made by multi cavity forging at RÁBA Axle Ltd. / Fig. 1/. The first step of multi cavity forging is upsetting. In the upsetting process the part is held by a robot and places it to the next working station. Upsetting must be inserted because of scale removal. When upsetting between parallel pressing plates the displacement field and the mathematically related wear can be modelled for the contacting part and the pressing plates [3]. This modelling requires the knowledge of the friction coefficient. Earlier we have tried to find out the friction coefficient from the largest diameter of the bulging part. In principle it is possible to find out the value of the friction coefficient from the diameter of the imprint appearing on the flat plate. When performing upsetting between parallel plates the contact area of the part and

the pressing plate (imprint) can be well observed, as during upsetting the smouldering steel causes a discoloration on the freshly prepared surface of the pressing plate. The imprint is only approximately circular. The diameter of the regression circle cannot be determined by traditional length measurement techniques during production.

The question arose whether it is possible to determine the diameter of the regression circle with acceptable precision without stopping the manufacturing process. In this work pictures made of a geometrical model are used instead of a real photograph and the precision of evaluation is studied on pictures of different qualities. This is partly done with the intention to understand better the capabilities of the available instrumentation and software and to extend their evaluation possibilities.

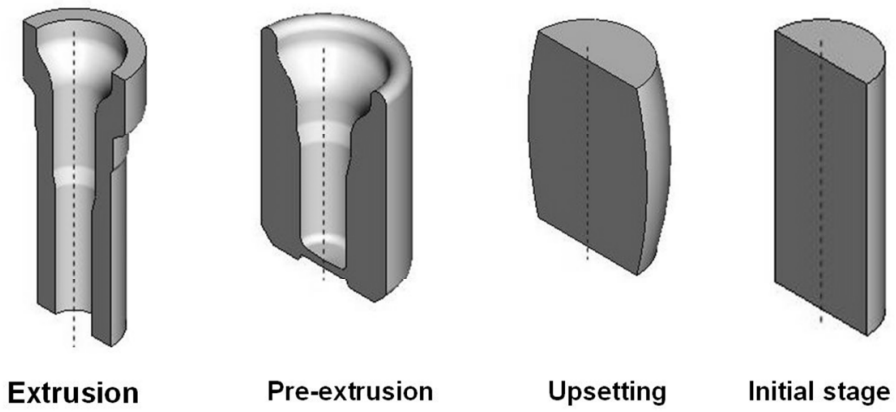


Figure 1. Intermediate stages of the axle stud

2. Principle of the evaluation

In order to achieve the goal one has to understand the basic principles of the evaluation, one has to develop a proper environment, including test possibilities, algorithms and software components that lead to the required results and are effective [4].

Basic principles are demonstrated using the MathCAD software [5]. Digital images consist of data points /pixels/. If the number of data points at the edge (perimeter) can be determined either by rows or by columns, knowing the pixel size the diameter of the regression circle fitting the limiting edge can also be determined. It can be shown that the precision of the evaluation depends on the size of the pixels /on the number of the pixels/. As a first approximation one can say that the more data points we have the higher the precision of the evaluation.

The possibility of the evaluation is demonstrated for the grey scale picture of a geometrical model of the pressing plate. The geometrical model of the pressing plate has been prepared by the Pro/Engineer CAD software /Fig. 2 /. It is easy to save pictures of various qualities (pixel number, hue, degree of compression) of the geometrical model then the effect of these parameters on the precision of the

evaluation can be studied. For the evaluation partly the MathCAD mathematical software, partly the software of an industrial CT was used.

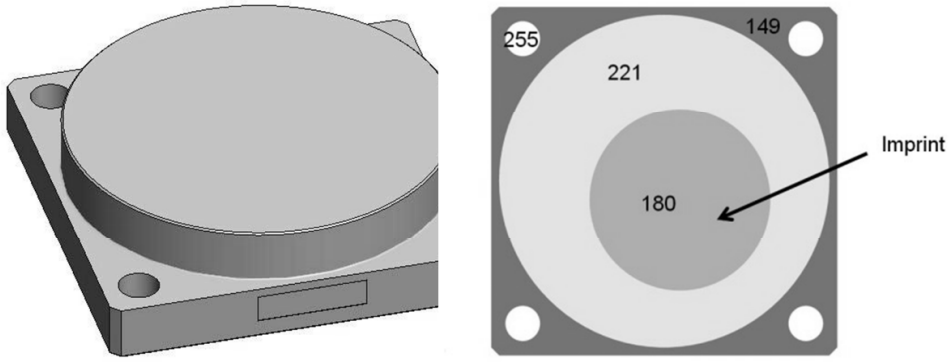


Figure 2. Geometrical model of the pressing plate, its top view together with the imprint and color coding

The MathCAD software stores the data points in a matrix of i x j size, where *i* the number of rows, *j* is that of the columns. Grey scale pictures make image processing easier, as only one colour code belongs to one data point. The value of the colour code varies between 0 and 255. 0 belongs to black, 255 to white. The code of the data points in the matrix can be changed. Where the number of pixels is to be determined, the code of the data points should be changed to 1, where not, it should be set to 0. In Fig. 2 the diameter of the circle coded as 221 is known / $D_0 = 210\,\text{mm}$ /, the diameter of the imprint /coded as 180/ is also known for the test exercise the control size is / $D_1 = 115$ mm/, otherwise the parameter to be determined is not known.

The result of the summed pixel numbers per column is an X column vector containing j elements, where the j^{th} element of the vector is the sum of the j^{th} column of the picture matrix. If plotting the sum of the column / X_i value / for each column of the picture matrix / at Y_j value / Fig. 3is obtained.

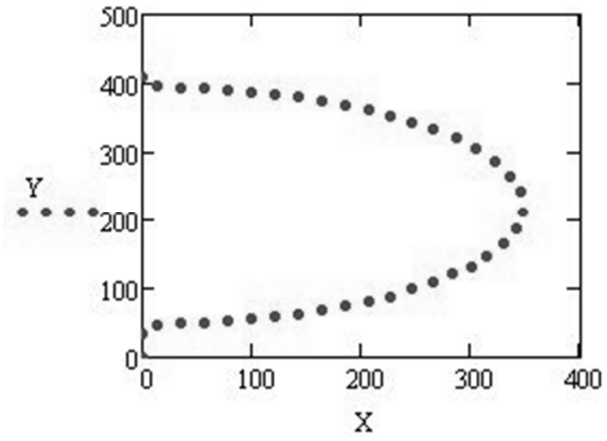


Figure 3. Distribution of pixel numbers within the imprint

One can see in Fig 3 that within the circle $X_j > 0$, while it is $X_j = 0$ outside of it. Wherever $X_j > 0$, the value of X_j in the position Y_j is proportional to the chord size at the given position.

If the X_j is divided by two, the obtained $X1_j$ and Y_j values can be regarded as coordinate values. Let $X2_j$ be the negative of the $X1_j$ values. Using the $X1_j$, $X2_j$ and Y_j values a symmetrical point set resembling a circle /Fig. 4 /.

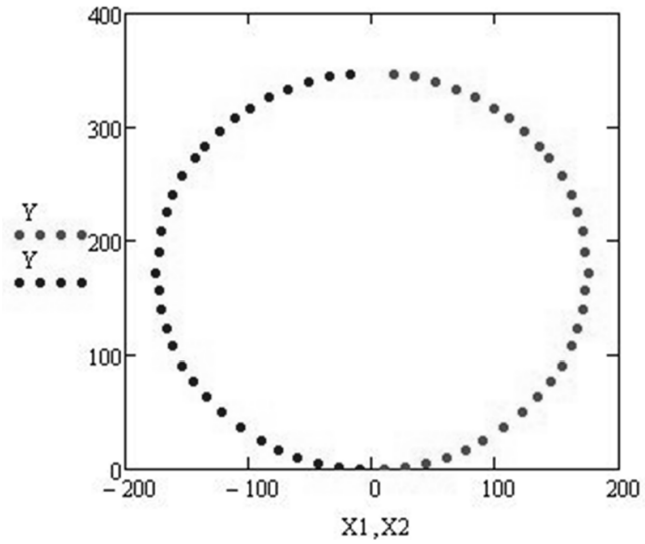


Figure 4. Plot of the circle obtained from the pixel numbers for the base circle

Without presuming a symmetric layout the evaluation can be performed if the figure is cut into two, roughly equal parts and the two halves are evaluated separately.

Using the MathCAD software a picture was evaluated, which is the image of the geometrical model made by the Pro/Engineer software.

In the case of the saved picture there were 1100 x 850 data points. In the case of the Pro/Engineer software it is possible to save the same geometrical model with resolutions of 200 dpi, 300 dpi, 400 dpi, 500 dpi and 600 dpi. Of course the pictures made with higher resolution contain more data points /see Table 1/ and of course the size of the saved file will also be larger.

Table 1. Data of the saved pictures

Serial number	Resolution [dpi]	Data points
1	100	1100 x 850
2	200	2200 x 1700
3	300	3300 x 2550
4	400	4400 x 3400
5	500	5500 x 4250
6	600	6600 x 5100

Pictures having 100 dpi resolutions are required to be processed by MathCAD software [6.].

The software of the industrial CT is capable of handling even pictures made with a resolution of 600 dpi /the evaluation possibilities will be shown later /.

Using Fig. 2 and the pixel number per column we get a point set /see Fig. 4/ which can be well approximated by a circle. In order to solve our task one has to know the parameters of the regression circle best fitting the point set /the coordinates of its centre, its diameter, D_0^* or radius R_0^*/. A possible solution for this is shown in the following chapter, here only the result is given and evaluated.

Expressing the radius of the base dimension by pixel numbers it is: $R_0^* = 318.699$ mm, $D_0^* = 637.398$ mm. As it is known, the original diameter D_0 was 210 mm. Using parameters D_0 and D_0^* a proportionality coefficient **c** can be defined:

$$c = \frac{D_0}{D_0^*} = \frac{210}{637.398} = 0.329464 \tag{3}$$

The diameter of the imprint can be expressed by pixel numbers as $D_1^* = 348.24$ mm. Knowing the proportionality coefficient the sought-for diameter of the imprint can be expressed as:

$$D_1 = c\, D_1^* = 0.329464 * 348.24 = 114.73 \text{ mm} \tag{4}$$

The D_1 diameter is known for the test exercise / $D_1 = 115$ mm /. It can be seen that the fitted value is inaccurate and one could improve it using MathCAD software as an evaluation tool only by significant program development, as the MathCAD software does support image processing, but there are several obstacles to its use:

- MathCAD software can process matrixes containing maximum 1 million elements (pixels) [6.],
- The exact determination of the border line of the imprint and the segmentation of the picture in some cases are very difficult.

A more precise result and a more effective image processing can be achieved by using the image processing software of an industrial CT. We will return to the precision requirement of the measurement.

3. Parameters of the regression circle best fitting the point set

In our work we first investigate the possibility of using the centre of gravity to determine the centre of the circle. We have not found reference to this in the literature. By centre of gravity we mean here the centre of gravity of the point set [7].

When speaking of the centre of gravity of a point set and if it can be assumed that weight of all points is uniform, the position vector of the centre of gravity of points with position vectors $a_1, \ldots\ldots, a_n$ is the arithmetic mean of the position vectors of the points [7]:

$$s = \frac{a_1 + \ldots\ldots + a_n}{n} \tag{5}$$

Thus the coordinates of the centre of gravity – which is supposedly identical with the centre of the circle – are as follows:

$$u0 = \frac{\sum_{i=1}^{n} x_i}{n} \qquad v0 = \frac{\sum_{i=1}^{n} y_i}{n} \tag{6}$$

When assessing the possibility of using the centre of gravity let us use an extremely simplified case with 3 data points. These three points define a triangle. In the case of triangles it is well known that the centre of the excircle and the centre of gravity are identical for a regular triangle.

Regular triangle /or polygon/ is obtained if the points are taken equidistantly along the perimeter of the circle. If it can be assumed that the edge points are the elements of the circle and are located equidistantly, then the centre of gravity of these points corresponds to the centre of the circle. In all other cases the centre of gravity can only approximately be regarded as the centre of the circle defined by the edge points.

Based on this the coordinates of the centre of a circle **u0, v0** can be obtained with good approximation if the centre of gravity of uniformly distributed points are taken.

Using the data points of Fig. 4 the calculated coordinates are **u0= 0; v0=318.5 mm.**

If the coordinates of the assumed circle (**u0** and **v0**) are known, the r_i radii can be calculated for the x_i , y_i point determined by the pixel numbers can be determined by distance calculation (7).

$$r_i = \sqrt{(x_i - u0)^2 + (y_i - v0)^2} \tag{7}$$

Calculating the r_i radii for the n data points one can determine that r radius which fits best the data points from the assumed centre.

It is accepted that the best approximation minimizes the squared sum of the deviations (8).

$$E(r) = \sum_{i=1}^{n} (r_i - r)^2 = \min \tag{8}$$

It can be proved that the minimum expression (8) can be obtained by the average of the radius values.

$$r0 = \frac{\sum_{i=1}^{n} r_i}{n} = 318.699 \cong 318.7 \text{ mm.} \tag{9}$$

The **r0** radius obtained from equ (9) gives the best approximation only from the original **u0, v0** centre. To get the complete solution a program was developed /Fig. 5 /, which varied the position of the original centre with uniform randomness, with Tx= ± 1 precision for N=10000 cases and the best R radius was recorded for each new centre together with the squared sum values of the deviations.

The data were stored in a matrix and the most advantageous of all was searched /Fig. 6 /. Random numbers were generated with a built-in function of MathCAD.

```
f3(MX,MY) := for k ∈ 0..1000
               u ← u0 + runif(1,-Tx,Tx)_0
               v ← v0 + runif(1,-Tx,Tx)_0
               for i ∈ 0..n - 1
                 x ← MX_i
                 y ← MY_i
                 SR_i ← sqrt((x-u)^2 + (y-v)^2)
                 SR
               r2 ← (∑SR) / (n + 1)
               E_k ← ∑(SR - r2)^2
               uu_k ← u
               rr_k ← r2
               vv_k ← v
               (uu vv E rr)
```

Fiure. 5. A MathCAD program to determine the parameters of the regression circle by the Monte-Carlo method

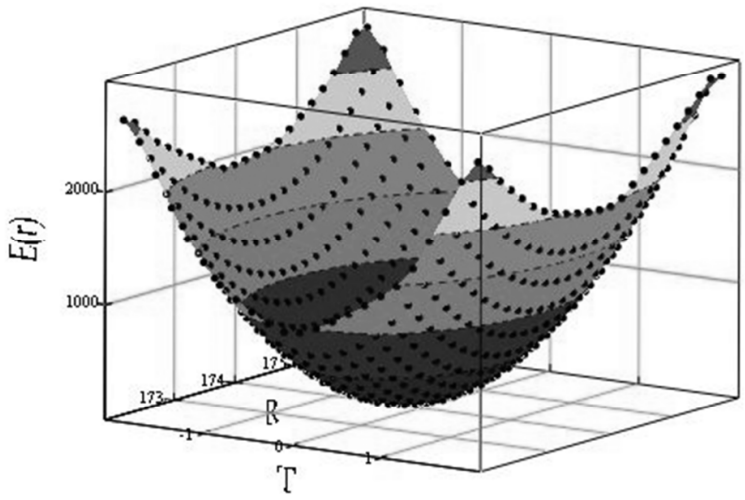


Figure 6. More precise determination of the initial parameters of the circle / r0, u0, v0 / by the Monte-Carlo method

The use of random numbers in solving mathematical problems is called Monte-Carlo method in the literature [8-9]. References [10-12] also use the Monte-Carlo method and least squares, but there three points are selected randomly and it is utilized that for a large number of points any randomly selected three points estimate well the parameters of the best regression circle.

The more precise value is $R_0^* = 318.699$ mm. u = -0.021; v = 318.476.

One can conclude that already the r0 value / r0 = 318.7 / gave a precise enough result for the given task for practical application. It means that if the pixel numbers of the picture of a circle are determined for each row or for each column within the circle, the centre and the radius of the circle defined by the pixel numbers can be well approximated by the (6, 9) relations.

Performing the evaluation for the imprint the diameter of the circle defined by the pixel numbers is 114.73 mm, the same as it was calculated using the equations (3, 4) in the previous chapter.

4. Use of the data processing software of an industrial CT

The data processing software of an industrial CT is Volume Graphics Studio (VGS). The main function of this software is the reconstruction of the data produced by the industrial Computer Tomography (CT). With its extra module this software package is capable of determining geometrical elements (planes, cylinders, circles etc.) and their parameters (size, position, orientation). The critical part of the evaluation is the segmentation of the pictures in this case too. Segmentation is performed by the VGS software in the given task using an edge detection algorithm based on the determination of greyness threshold: a certain level of greyness is assumed which corresponds to the

transition from one material to the other or to the environment. This threshold can be rarely defined by a concrete greyness value therefore the software calculates also sub-pixels by interpolation. The definition of the threshold is made by using a histogram. When plotting the greyness values and the corresponding pixel numbers, various materials (in the picture areas various colours) yield peaks on the histogram. In the industrial CT applications it is an accepted practice is to use the so-called ISO-50% threshold value. This value is the geometric mean of the greyness values belonging to the two peaks. According to our experience the treatment of the histogram involves a certain degree of subjectivity, which influences somewhat the precision of the evaluation / see Fig. 8 /.

One can fit a geometrical element onto the points of the limiting curve by the least square method.

Pictures of different resolutions of the test exercise / see Table 1 / are accepted by the highly developed processing software of the industrial CT without any problem.

When processing the saved pictures the transition zone is the critical part. In this transition zone the grayness value of the pixels varies, in the transition zone smaller areas of smaller size / 16-20 pixels / can be observed, almost independently of the resolution /Fig. 7 /. Of course the size, portion of the transition zone is smaller on pictures of higher resolution / more pixels /.

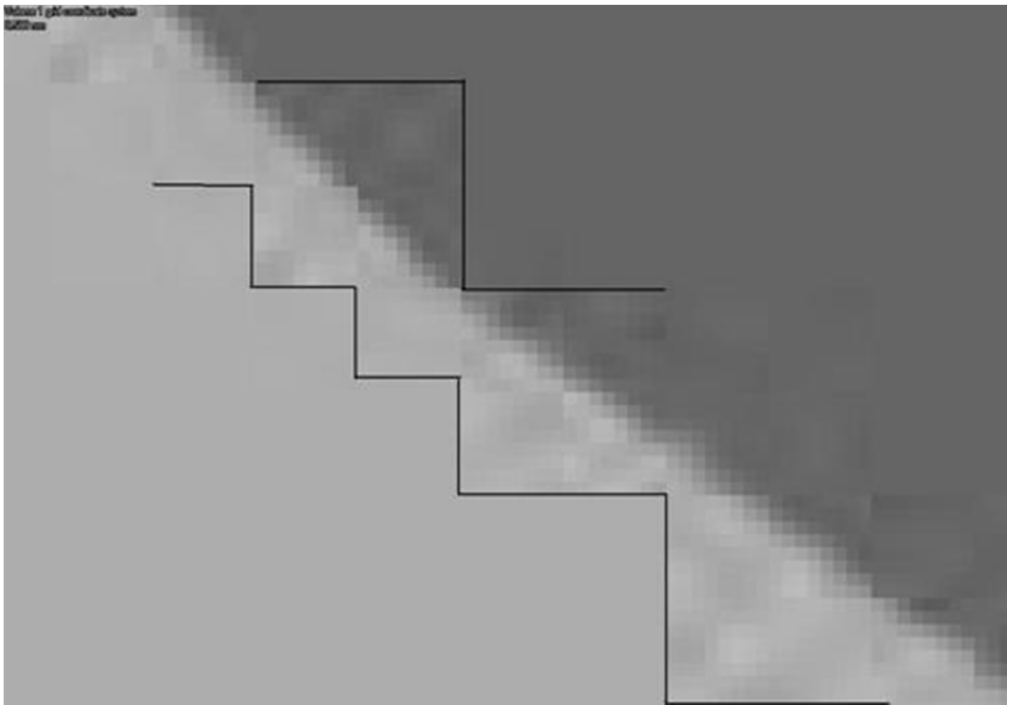


Figure 7. The transition zone with parts of varying greyness

When processing a retrieved picture the size of single pixel should be defined in advance. If the pixel size is the unity, the result - the diameter of the regression circle – will be expressed by the pixel numbers. Accordingly the calculation of the searched size is very similar to the MathCAD application. The results obtained are summarized in Table 2:

Table 2. Evaluation using the software on an industrial CT

Serial number	ISO 50%		pixel size	Diameter [mm]	
	reference	imprint	[mm]	reference	imprint
1	224	168	0,294308	210,000	114,982
2	224	168	0,147154	209,998	114,986
3	224	169	0,098104	210,000	114,998
4	222	159	0,073558	210,000	114,990
5	222	159	0,058851	210,002	115,000
6	222	159	0,049045	210,000	115,006

Based on the data of Table 2 one can see that with increasing pixel numbers the precision of the evaluation can be improved to a certain degree, but above a certain pixel number (resolution) the improvement is not clear. It can be well observed by plotting the results graphically /Fig. 8 /.

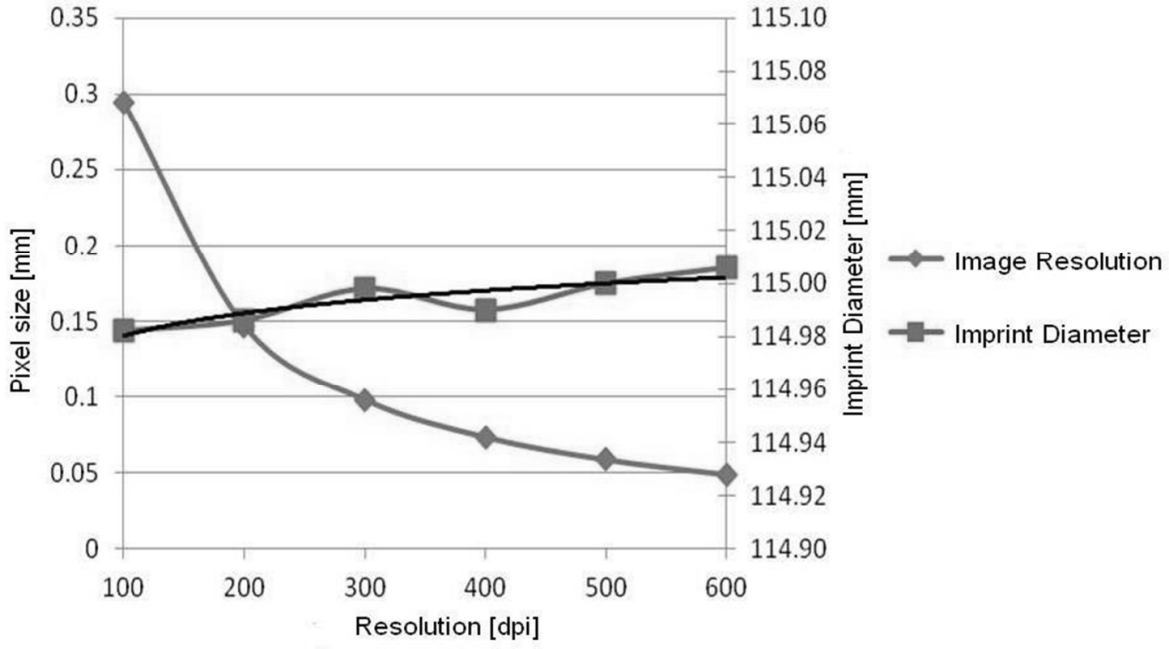
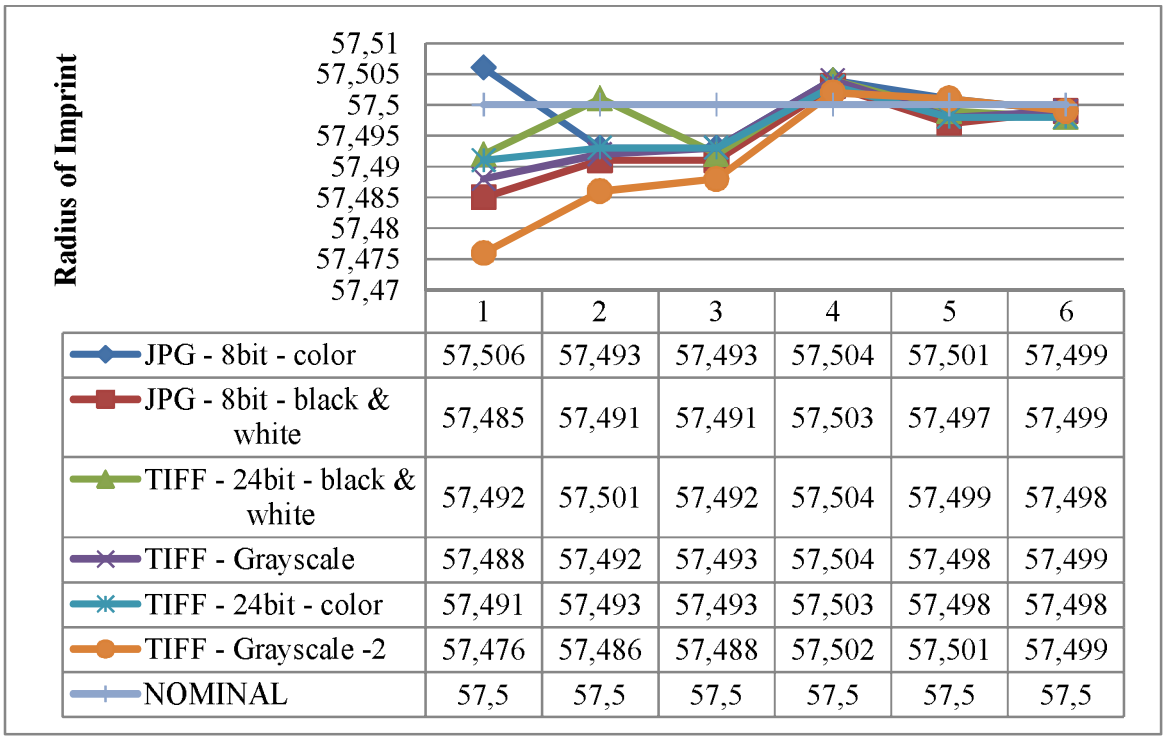


Figure 8. The effect of pixel size (resolution) on the determined size /imprint diameter/

In addition to the pixel numbers an important characteristic of the picture is the hue and how these two parameters are zipped in the picture data file. We have investigated how these variables affect the precision of the reconstruction. We have saved the picture of the test exercise into two well-known file formats, namely into the zipped JPG format and into the complete TIFF format. The first is advantageous because of its smaller size, while in the latter the quality of the picture is not degraded. Hue value was ascribed to the picture points at two levels, the number of test values was increased by 256 colour (8 bit) and 16,7 M colour (24 bit) variants. According to the results presented in Fig. 9

the best approximation can be achieved by increasing the resolution, hue values and the zooming produced differences only at low dpi values.



	1	2	3	4	5	6
JPG - 8bit - color	57,506	57,493	57,493	57,504	57,501	57,499
JPG - 8bit - black & white	57,485	57,491	57,491	57,503	57,497	57,499
TIFF - 24bit - black & white	57,492	57,501	57,492	57,504	57,499	57,498
TIFF - Grayscale	57,488	57,492	57,493	57,504	57,498	57,499
TIFF - 24bit - color	57,491	57,493	57,493	57,503	57,498	57,498
TIFF - Grayscale -2	57,476	57,486	57,488	57,502	57,501	57,499
NOMINAL	57,5	57,5	57,5	57,5	57,5	57,5

Figure 9. Measured values obtained on pictures of different hue values, degree of zipping and resolution and their relation to the nominal value

The precision of 0.1 mm corresponding to the calliper could not be achieved by using the MathCAD software, however, the industrial CT clearly allowed achieving the required precision.

It can be concluded that by the VGS software the diameter of the imprint can be determined much more precisely than by the MathCAD software. The VGS software is user friendly, e.g. it recognizes the circular shape if 3 points are identified. More evaluations are possible on the same picture; one does not have to change the picture only the assignment has to make systematically.

5. Conclusions and Future Improvements

If image processing is to be used to determine an unknown geometrical parameter /diameter/ the effects of pixel number, hue value and degree of zooming can be well studied by a geometrical model. Using this geometrical model digitalized pictures of different qualities can be easily saved at constant settings. According to the results of the test exercise the precision of the evaluation is most influenced by the resolution of the picture. Hue levels and the degree of zooming of the saved picture affect much less the geometrical size.

Based on the evaluation of the test exercise it can be concluded that the determination of the imprint diameter using a digital photograph cannot be performed using the MathCAD software with the required precision and efficiency under the given

conditions, while the processing software of an industrial CT can be used with a high level of technical safety.

Using an example it has been demonstrated that determining the pixel numbers within a circle for each column or row of a grayscale picture of a circle the centre of the circle deduced from the pixel numbers can be well approximated by the centre of gravity of the point set.

6. Acknowledgment

This question was originally raised by Ferenc Tancsics, a PhD student, now leader of forging technology at Rába Axle Ltd. We acknowledge his permission to study this problem independently. We thank for the management of Rába Axle Ltd. allowing the investigation of the pressing plate at the Széchenyi István University. In return we provide the results of the research.

The research work presented in this paper was carried out as part of the TÁMOP-4.2.2.A-11/1/KONV-2012-0029 project in the framework of the New Széchenyi Plan. The realization of this project is supported by the European Union, and co-financed by the European Social Fund.

References

[1] Dr. Lakatos István, Titrik Ádám, Orbán Tamás: Belső égésű motor modell felállításhoz szükséges adatok meghatározása (*Data determination of an internal combustion engine for model set-up (in Hungarian))*, IFFK 2011, Budapest, Hungary, Magyar Mérnökakadémia, pp. 151-157 (2011)

[2] Dr. Lakatos István, Titrik Ádám, Orbán Tamás: *Data determination of an internal combustion engine for model set-up*, HUNGARIAN JOURNAL OF INDUSTRIAL CHEMISTRY 39**:** 35-40 (2011)

[3] Ernő Halbritter; Ferenc Tancsics: *Melegalakító kovácsszerszámok kopásvizsgálata (Study of wear of hot forging tools, in Hungarian)*, A jövő járműve, 1/2: 29-35 (2012)

[4] Ferenc Firtha: *Trikromatikus és hiperspektrális képfeldolgozási módszerek élelmiszerek és termények vizsgálatára* (Trichromatic and hyper spectral image processing methods for studying foodstuff and produces, in Hungarian), PhD Thesis, Department of Physics and Automatics, Faculty of Food Science, Corvinus University of Budapest, Budapest (2008)

[5] Mark Nixon; Alberto S Aguado: *Feature Extraction & Image Processing, Second Edition*, Academic Press is an imprint of Elsevier (2008)

[6] *MathCAD Help*, PTC, Mathcad 14.0

[7] György Hajós: *Bevezetés a geometriába (Introduction to Geometry, in Hungarian)*, Tankönyvkiadó, Budapest (1962)

[8] József Cserti: *A munkára fogott véletlen I. (Chance exploited, I., in Hungarian) Középiskolai Matematikai és Fizikai Lapok*, 53/7: 432-436 (2003)

[9] József Cserti: *A munkára fogott véletlen II. (Chance exploited, II., in Hungarian) Középiskolai Matematikai és Fizikai Lapok*, 53/8: 493-497 (2003)

[10] Bálint Laczik: *Regressziós kör illesztése adott síkbeli pontsokasághoz (Fitting a regression circle to a given point set in a plane, in Hungarian)*, A Dunaújvárosi Főiskola Közleményei; XXX. 1. (2008)

[11] Bálint Laczik: *Kör-regresszió kombinált sztochasztikus-analitikus módszerrel (Circle regression by combined stochastic-analytical method, in Hungarian)*, in Proceedings of VI. Gépipari Méréstechnikai Tanácskozás (in Proceedings of VIth Meeting on Mechanical Measurement Technology, Győr), October 20-21 (1986)

[12] Gonzales, R. D., Woods, R.E.,: *Digital Image Processing*, Prentice Hall (2007)

New analytical method for engine diagnostics based on pressure indication of cylinder clearance

Dr. Lakatos István Ph.D.

Széchenyi István University, Department of Automotive and Railway Engineering,
Egyetem tér 1., 9026 Győr, HUNGARY
E-mail: lakatos@sze.hu

Abstract: Engine indication is a measuring method applied primarily in the field of engine development. The development of sensors have made it possible to apply it - without dismantling - for diagnostic methods.

We have elaborated its methods within the frame of a research project. We are to introduce the results of this work in the forthcoming study.

Keywords: indicated pressure, mean indicated pressure, indicating spark-plug,

1. Introduction

With the traditional analytical methods only the engine performance on the crankshaft could be measured (P_e, namely effective performance). The pressure change of the process which is taking place in the cylinder can be determined by measuring the pressure of the cylinder clearance or as it is defined by the technical literature with indication.

In internal combustion engines the chemical energy of the fuel taken in within the given cycle transforms into thermal energy, then - due to the change of pressure and volume - into power. In the meantime the volume of work space periodically changes. **During the work procedure the pressure of the working media constantly changes and this pressure change can be shown in the form of an indication diagram.**

The indication diagram displayed according to the piston shift of the four–stroke engine can be seen in Figure 1., while the diagram according to the angular displacement can be seen in Figure 2. In view of the features of crank mechanism, one of the diagrams contributes to the design of the other.

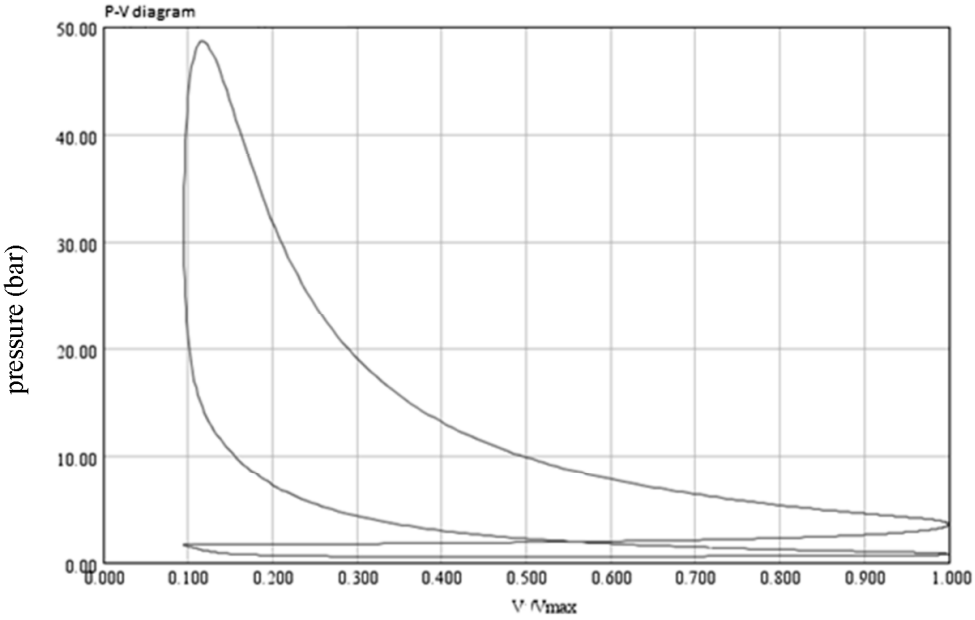


Figure 1. indication diagram displayed according to the piston shift

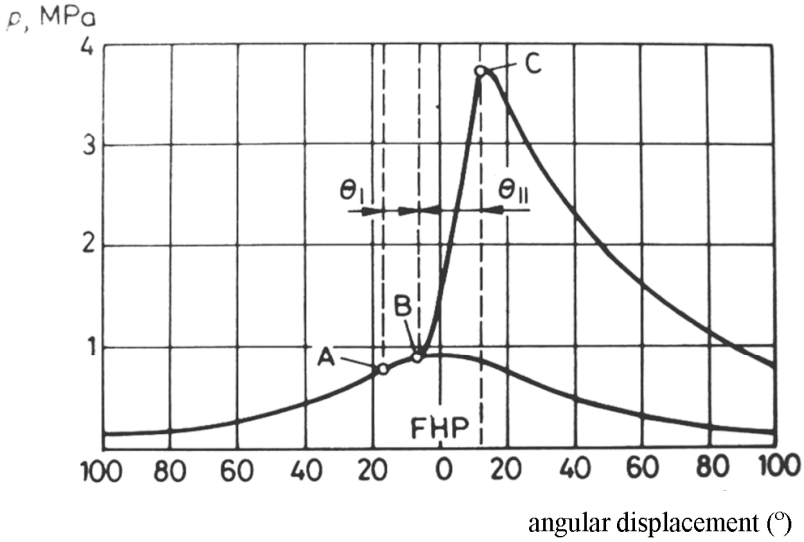


Figure 2. indication diagram displayed according to angular displacement (A – ignition, B – the beginning of the combustion, C – maximum pressure, Θ_I – ignition delay, Θ_{II} – pressure increase section)

2. Indicated features

The mean indicated pressure is the mean height of the efficient space of p-V indicator diagram. **The efficient space of p-V indicator diagram defines the value of the indicated work for a cycle**. If we divide the indicated work with the cylinder displacement we get the mean indicated pressure:

$$p_i = \frac{w_i}{V_H} \qquad \left[J/m^3 \qquad vagy \qquad N/m^2\right]$$

(The mean indicated pressure is expressed in kPa or MPa.) It is obvious from the measuring units that the mean indicated pressure can be defined as well as the indicated work derived from unit cylinder displacement.

In view of the indication diagram the indicated work can be defined with clearance measurement. The clearance of the charge change process of uncharged two- or four- stroke engines has negative sign.

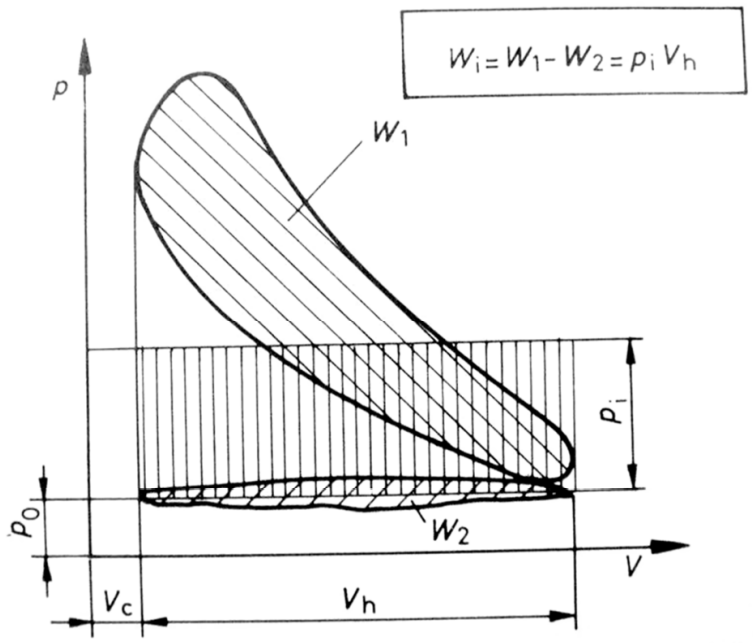


Figure 3. Indicated features

The indicated mean pressure changes along with the load of the engine, it reaches its lowest value in the idling position of the engine, here the indicated work only meet the energy needs of friction and auxiliary equipment (p_i=p_m).

Indicated work calculated from the indication diagram of one cylinder of the engine during one work cycle

$$W_i = p_i \cdot V_H \qquad [Nm]$$

Where:

- p_i the mean indicated pressure
- V_H the cylinder displacement.

The number of work cycles per seconds $\frac{2n}{i}$, where **n** is the revolution number of the crankshaft in 1/s; **2n** is the number of piston stroke per seconds, **i** is the number of strokes during one work cycle. On the basis of this the indicated performance of the engine with **z** number of cylinders:

$$p_i = \frac{2}{i} \cdot p_i \cdot z \cdot V_H \cdot n \qquad [W]$$

Where:

- in case of four-stroke engine: i=4,
- in case of two-stroke engine: i= 2.

The performance is generally expressed in kW. The indicated power efficiency and the indicated specific fuel consumption are the economic efficiency indices of the indicated features of the engine.

The unit used for specific fuel consumption in practise is **g/kWh**. In case of natural gas vehicles the specific fuel consumption is defined according to volume unit, **m^3/kWh**, or **MJ/kWh** specific thermal energy consumption is applied to describe the efficiency of the engine.

The measured performance on the crankshaft of the engine, the effective performance, is lower than the indicated performance calculated on the basis of the indication diagram. A definite proportion of the indicated performance is consumed to move the components of the engine shifting on each other and to sustain the continuity of the operation of the engine. By signing this necessary loss with P_m the effective performance of the engine can be expressed in the following way:

$$p_e = p_i - p_m$$

The notion of mean effective pressure (p_e) and the mean pressure (p_m) featuring the mechanical losses can be defined by the mechanical performance loss and the work of effective performance for a work unit- as in case of indicated work- according to a unit displacement. With the help of them the adequate performances can be expressed like in case of the indicated features.

The mechanical losses can be expressed even with the help of the mechanical power efficiency:

$$\eta_m = \frac{P_i - P_m}{P_i} = 1 - \frac{P_m}{P_i} = \frac{P_e}{P_i}$$

$$\eta_m = \frac{p_i - p_m}{p_i} = 1 - \frac{p_m}{p_i} = \frac{p_e}{p_i}$$

One part of the mechanical losses of the engine is caused by friction (firstly between pistons and cylinders, secondly bearing frictions) which is signed by P_s. We mean by the performance needs (P_b) of the auxiliary equipment the needs of the oil pump, the water pump, the ignition or injection device, the fuel pump and depending on the standards about recording the measuring conditions the performance need of the cooling fan. In addition the performance need of the compressor (P_k) in case of two-stroke engines without crank chamber compression (and mechanically charged four-stroke engines). Thus the relation between effective and indicated performance is the following:

$$P_e = \eta_m P_i$$

3. DIAGNOSTIC METHOD BASED ON ENGINE INDICATION

3.1. Analysed vehicle and engine

The vehicle selected for the analysis was a VW Jetta (Figure 4.).

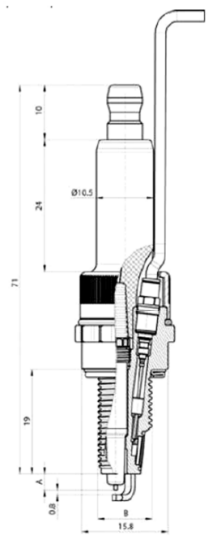
Figure 4.: The analysed vehicle

3.2. The compiled measurement system

Engine indication is a fundamental method for engine experiments. In the project we took up the challenge to elaborate a new diagnostic method (analysis without or with partial disassembly).

For measuring we chose the following type of pressure gauge spark-plug (according to the dimension and the heat range of the spark-plug prescribed for the engine):

GG1452 INDICATING SET GH13Z-24/ ZF43 F7L PRT



Other dimensions
A=The position of the electrode

Figure 5: Indicator spark-plug

AVL DPM-800 spark-plug for checking cylinder clearance completely contains the pressure gauge integrated into the spark-plug.

The newly developed indicator spark-plug is suitable for very precise pressure measurement without disturbing or influencing the combustion procedures anyhow. Available with M10, M12 bore diameters, different length and 3, 5 and 7 heat range.

Built-in detecting piezocrystall, namely GaPO4 (gallium-phosphate) unit sensitivity 12 pC/bar. The new device can provide aid not only in engine development, in chip-tuning, but in the field of engine diagnostics of repair industry.

The primary objective of the development of the pressure gauge of the spark-plug was to achieve an adequately long lifespan during test circumstances. The platinum-electrode is such a component of the spark-plug which has adequate solidity and bears thermal strains for long period. The body electrode is also platinum-tipped. During the 30.000 km long test cycle no failure was experienced by the developers in the combustion and pressure gauge function of the spark-plug. Due to the transparent modular structure even the end user can replace the components of the units.

The construction of the measuring system:

The new measuring system was designed by using the following components:

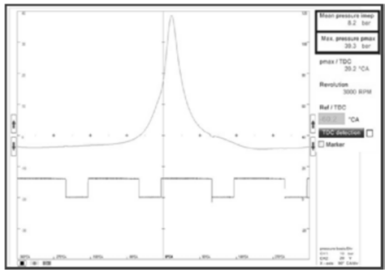
AVL piezo spark-plug	
AVL amplifier unit	
AVL DiScope 802 dual channel oscilloscope	
AVL DiX modular measurement software	

Table 1.: components of the new diagnostic system for measuring pressure

The installation of the indicator spark-plug had to be conducted with due foresight.

3.3. Analysing working points

The load feature range of combustion engines consists infinite number of working points.

For the purpose of our analysis we have selected some of these points located on the characteristic curve. The working points of the characteristic curve are stationary working points, which mean that the characteristic features of the engine (revolution number and load) defining the working point are kept at a constant level. The regulated operation of today's engines affects the cycle stationarity within the working points. This fact will be analysed later.

To serve the purpose of the analysis we have chosen two types of characteristic curves (Figure 6.):

- the rolling resistance characteristic curve (Fv ~ v2)

  1. load characteristic curve,

- the external (complete load) moment characteristic curve

  2. load characteristic curve.

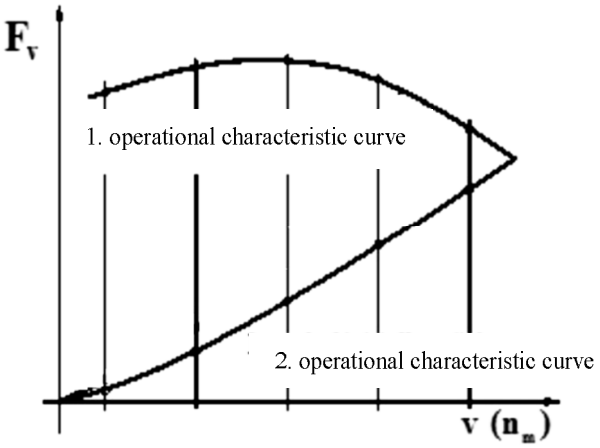


Figure 6.: analysed operational characteristic curves

In the following part of the study we summarize the values of indicator diagrams at the given measurement points of the characteristic curve. The values in the rows of the diagrams are the data of the indicator diagram characterising the work cycle of the examined cylinder of the engine. The deviation of the values derives from the fact that variable work cycles are added to the average (constant) level of the performance achieved through the wheels due to the operation of the engine. This variable value is strengthened by the control cycles located on the engine, such as the regulation of lambda or combustion knock. Their operation cause slight continuous changes on operating parameters.

The working point values are in every case the average of the cycle values. Some photos are shown below to illustrate the measurements.

Figure 7.: installation of the indicator spark-plug

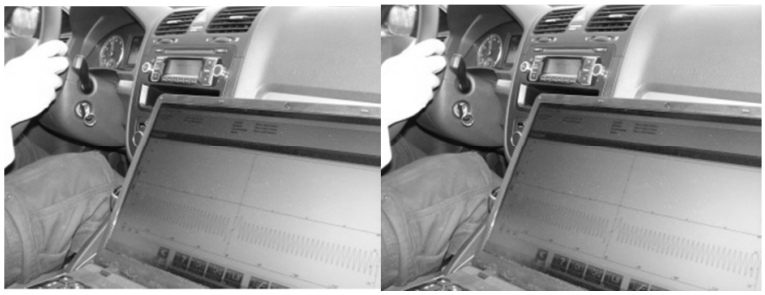


Figure 8.: Measurement on the motorway

The majority of the measurements due to reproductivity were accomplished in the Vehicle Diagnostic Laboratory of the Department of Automotive and Railway Engineering on free rollers.

The measurements were accomplished even in flawless conditions and with simulated failures.

Velocity 60 km/h; load 100%						
	Mean indicated Pressure		**Maximum combustion pressure**		**pmax/FHP**	**Revolution number**
	bar	**N/mm^2**	**bar**	**N/mm2**	**degree**	**Rev/min**
1.	4,3	43	14,9	149	112	2280
2.	4,5	45	16,3	163	110,7	2270
3.	4,5	45	17,1	171	111,3	2270
4.	..	..	..	..	..	..
19.	4,7	47	16,4	164	110,7	2230
20.	4,5	45	22,5	225	109,9	2230
Max	4,7	47	22,5	225	114,8	2280
Min	3,7	37	12,1	121	109,8	2220
Average	4,29	42,85	15,33	153,25	112,02	2244,50
Difference	1	10	10,4	104	5	60
Variation	0,22	2,23	2,11	21,13	1,50	20,38

Table 2.: cycle features of square load characteristics (v=60 km/h)

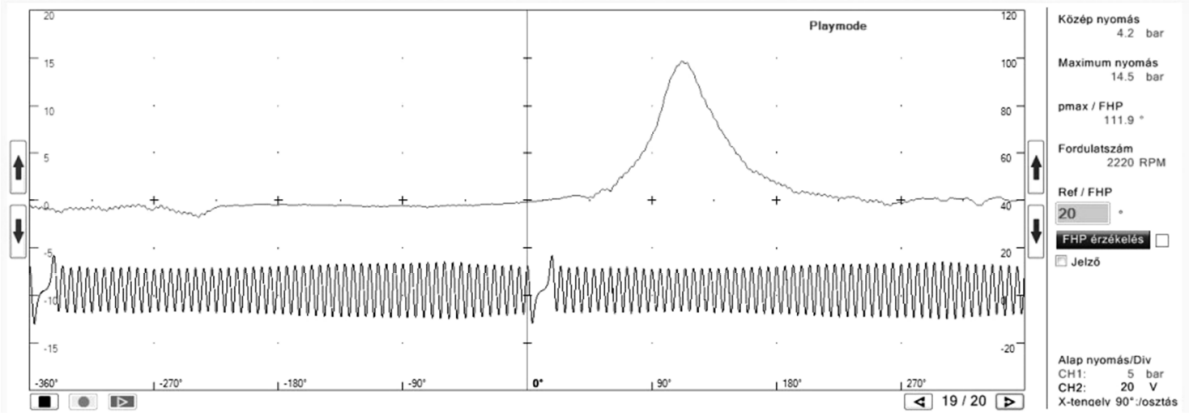


Figure 9.: indicator diagram (Working point: square load characteristics, v=60 km/h)

After the flawless conditions let's have a look at some simulated failures, or more precisely their consequences on diagrams and measured registers:

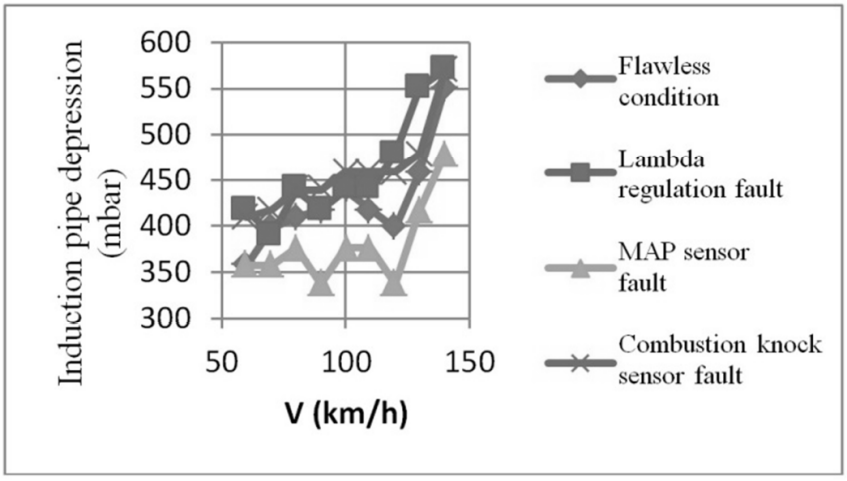


Figure 10.: failures measured along with square load characteristics compared in the view of induction pipe depression

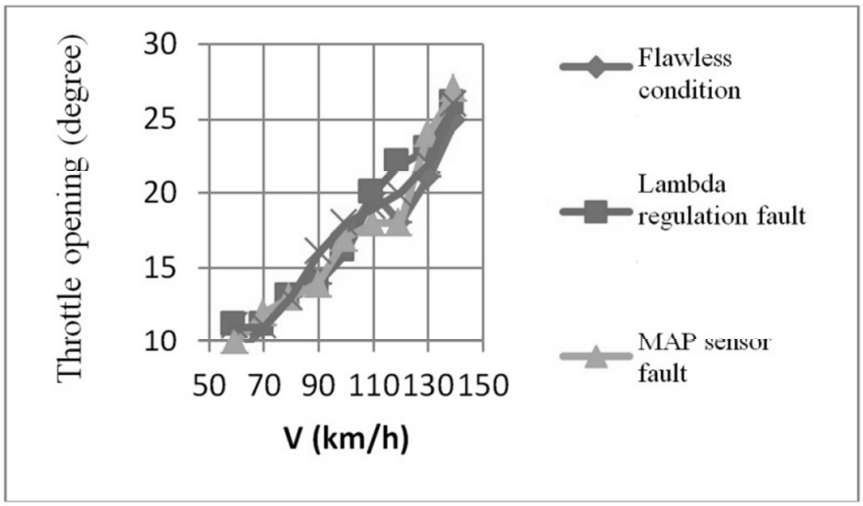


Figure 11.: comparison of failures measured along with square load characteristics in the view of opening angle of the throttle

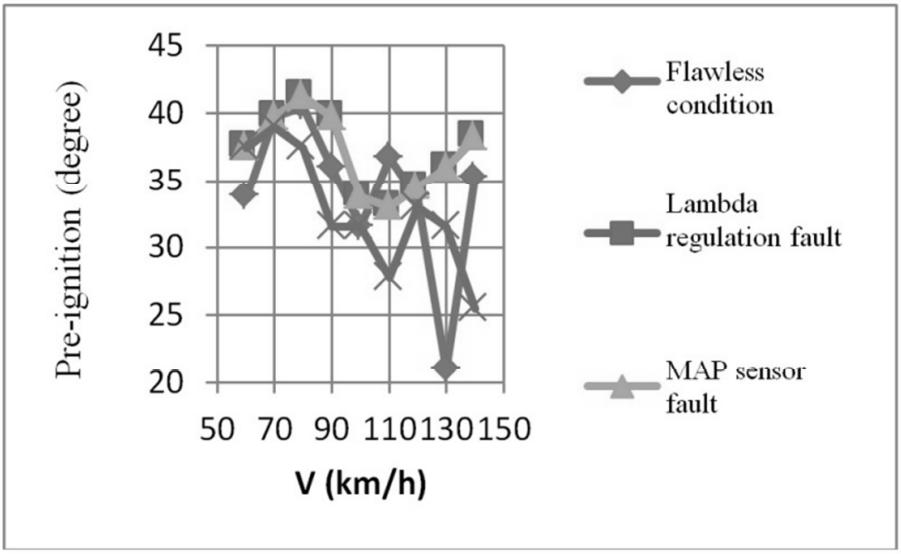


Figure 12.: comparison of failures measured along with square load characteristics in the view of opening angle of the pre-ignition

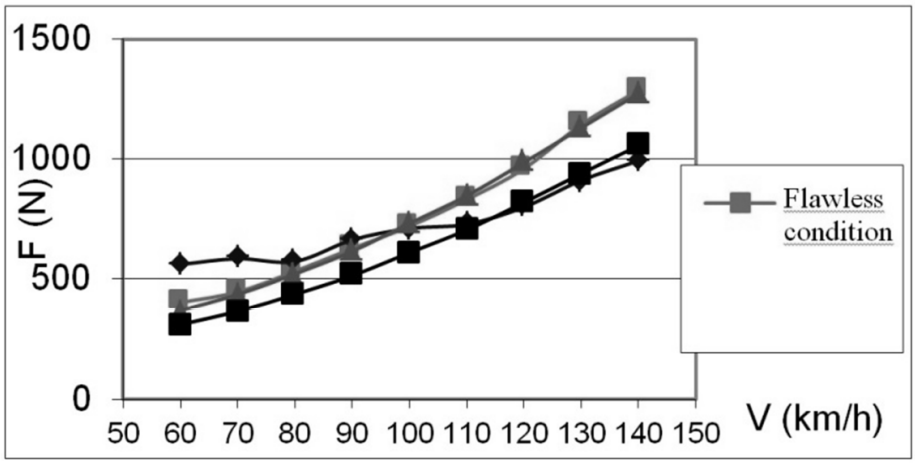


Figure 13.: comparison of failures measured along with square load characteristics in the view of tractive force

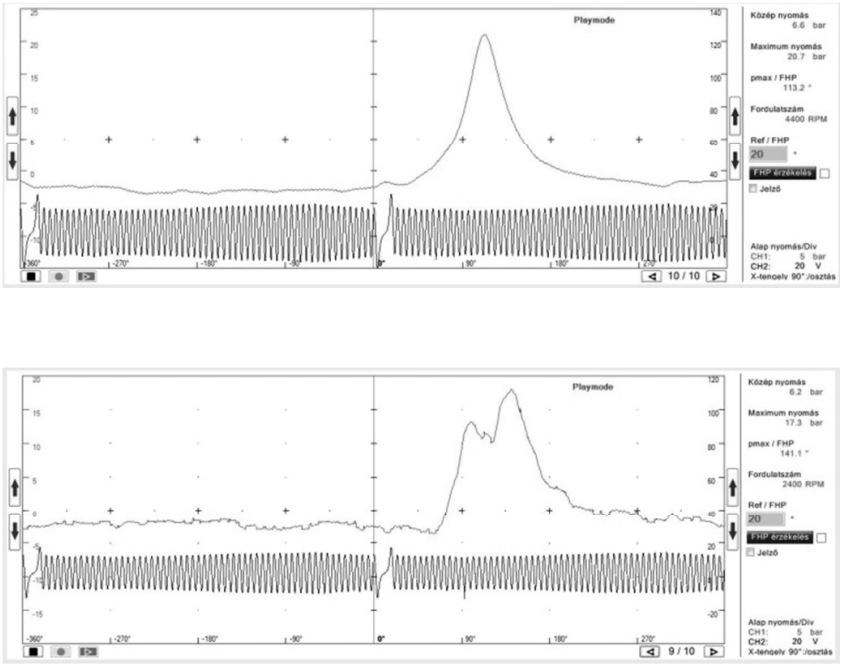


Figure 14.: Combustion knock sensor fault (maximum load): the picture above is flawless, the one below is faulty

4. CONCLUSIONS, ASSESSMENT

In case of Otto engines the variation of work cycles is typical on given working points. The reasons derive from the values differ in time and space of combustion velocity. This is caused by the fact that at different points of the combustion chamber the mixture formulation is not homogeneous. With the improvement of the quality of the mixture formulation the range of cycle variation can be reduced.

The average of the cycle features differing form each other is typical to the given working points. Due to this in the tables recorded at the working points with measurements we have shown the values of the average and the deviation.

The variation of the cycles is influenced by the controlled systems of the engine, too, as at the given working points they affect the setting features of the engine. Thus, if the controlled systems are excluded (e.g.: lambda- and combustion knock control) the rate of cycle variation is reduced. Concerning this, several measurements have been accomplished in the previous chapter.

The biggest recognizable difference triggered with the exclusion of combustion knock sensor, in both of the cases of maximum and square load. In this case the maximum load condition is more characteristic since the risk of combustion knock is the highest here.

The newly designed measurement system and method make it possible to apply engine indication, which was only used formerly for research and development, for diagnostic purposes.

Its significance is that the introduced procedures make it possible to conduct measures even on the highways due to the flexibility of the formulation of the measurement system.

For measures we have to possess reference pictures and data about the analysed type which can be obtained by measuring reference vehicles with flawless technical conditions. We should assign the reference data to the measured operation identification features.

During the further diagnostic analyses, we measure at the same working points as of the reference state, and the measured pictures, data are to be compared with the features of the given type recorded at its flawless state.

We should take into account the cycle variations and always take the average picture as a basis from the cycle pictures of measured time intervals (e.g. 20s).

The advantages of the laborscope integrated into the measuring instrument can be utilized at the examination to accomplish the oscilloscope analysis of sensors and interferers besides pressure analysis. For this, reference pictures are stored for the measuring notebook, moreover there is availability to store more data on it.

The prototype of the developed new measurement method and measurement system within the scope of the research is such a practically applicable digital measuring unit or measurement system which is not only adequate for indicating the cylinder clearance of the engine (measuring pressure) but with the help of the integrated digital oscilloscope the signals and characteristic features of electronic transmitters can be displayed and compared with reference pictures or with one another.

During the research work we tested the tool for weeks, gathered information and measurement results.

On the basis of this according to its usage the measurement system can be divided into 2 main types:

Direct measurement: when one of the signs of the transmitter of the vehicle is displayed on the screen, and by considering the value and shape of this electrical sign we can deduce to the failure. In case of e.g.: generator, ignition system, transmitters, etc.

Indirect measurement: We record reference values with the instrument and we compare them with the later measured values or signs (E.g.: pressure analysis of the cylinder clearance or even the analysis of transmitters, interferers can be accomplished).

The process of indirect measurement:

Recording reference signal or value

The reference signal is actually the comparative parameter of the original state (Those service-stations which have availability to the original parameters of the vehicles serviced by them do not have to take reference data, just obtain these original data)

The database of reference signal can be voluntarily extended:

The recording of the reference signal can take place on straight, level road while the vehicle is moving (level road rolling resistance load or rolling road).

According to our experiences the speed at 100 km/h is suitable for that. It is advisable to look for a relatively long straight flat section of a motorway. Be careful, as a hill- or slope running or load of the vehicle may significantly affect the measurement. The instrument can record altogether 100 pictures during a measurement. It is suggested to be utilized by the users namely the length of recording signs should last at least 1 minute.

Recording signs of diagnosed/faulty vehicle

The process of recording signs of faulty vehicle takes place exactly in the same way as it was described at the reference signs. The process of evaluation is also equivalent.

Revealing the fault

The fault is revealed by comparing the reference and the faulty sign. Here we have to pay attention to the values and the shape of the sign.

With the help of the comparison of the two figures it can be unambiguously stated that there is a significant variance both between the sign shape and the values of maximum and mean pressure.

The practise of everyday application will provide the routine of decision making in diagnostic measurement for which the storage of more faulty and reference pictures in the database is needed.

References

[1] Dr. Lakatos István, Titrik Ádám, Orbán Tamás: Belső égésű motor modell felállításhoz szükséges adatok meghatározása (*Data determination of an internal combustion engine for model set-up (in Hungarian))*, IFFK 2011, Budapest, Hungary, Magyar Mérnökakadémia, pp. 151-157 (2011)

[2] Lakatos István: The effects of charge change timing on the operation of uncharged Otto engines, Ph.D. dissertation, BME, 2002, 112 p.

[3] Dr. Lakatos István: Optimalisation of the charge replace process of uncharged Otto engines of OHC control, Hungarian Electronic Journal, Győr, under construction

[4] Dr. Lakatos István: Untersuchung der Zusammenhängen zwischen der indizierten Werten und der mit Rollenprüfstand gemessenen Versuchsergebnissen, Járművek, 2002

[5] AVL DISCOPE 802 GÉPKÖNYV, a 1.6.0.192 program verzióból, AVL DITEST FAHRZEUGDIAGNOSE GMBHALTE POSTSTRASSE 152A-8020 GRAZ, AUSTRIA, 2011

[6] AVL DITEST DPM 800 gépkönyv, AVL DITEST FAHRZEUGDIAGNOSE GMBHALTE POSTSTRASSE 152A-8020 GRAZ, AUSTRIA, 2011

[7] Benjamin Robert Brown: Combustion Data Acquisition and Analysis, Loughborogh University, Department of Aeronautical and Automotive Engineering

[8] Dipl.-Ing. Gerald Rämisch: Modellbasierte Diagnose am Beispiel der Zylinderdrucksensorik von Ottomotoren, Isenbüttel, 2009

[9] Verbrennungsmotoren / Prof. Dr. Jan Czerwinski / Assistent Dipl. Ing. Thomas Hilfiker / in Zusammenarbeit mit KISTLER Instrumente AG – Experten: Hans-Jörg Gisler / Christian Bach: Verbrennungsdiagnostik mittels Druckindizierung

[10] Mark C. Sellnau Delphi Central Research and Development Frederic A. Matekunas, Paul A. Battiston and Chen-Fang Chang General Motors Research and Development Center David R. Lancaster General Motors Powertrain Group: Cylinder-Pressure-Based Engine Control Using Pressure-Ratio-Management and Low-Cost Non-Intrusive Cylinder Pressure Sensors, SAE TECHNICAL PAPER SERIES, 2000-01-0932

[11] Josef Blažek: THE COMBUSTION PROCESS ANALYSIS BY MEANS OF IN-CYLINDER PRESSURE MEASUREMENT, Međunarodni naučni simpozijum Motorna Vozila i Motori International Scientific Meeting Motor Vehicles & Engines Kragujevac, 04. - 06.10.2004

[12] RAINER MÜ LLER, HANS-HUBERT HEMBERGER, and KARLHEINZ BAIER Daimler Benz AG, Research Institute 1, F1M/EA, HPC T721; 70546 Stuttgart, Germany: Engine Control using Neural Networks: A New Method in Engine Management Systems, Meccanica 32: 423–430, 1997., © 1997 Kluwer Academic Publishers. Printed in the Netherlands.

Omnidirectional Wheel Simulation – a Practical Approach

Viktor Kálmán

Budapest University of Technology and Economics,
Department of Control Engineering and Information Technology,
Magyar Tudósok körútja 2., Budapest 1117
Phone: 1 463 4025, fax: 1 463 2204
e-mail: kalman@iit.bme.hu

Abstract: In this paper a configurable omnidirectional wheel model is presented, which can be used for dynamic simulation and parameter tuning of omnidirectional robotic systems. First a brief overview is given on well known omnidirectional wheel designs and models from the literature, then two modeling approaches are described and compared. The usability of the models is verified by simulation, using two omnidirectional platforms and kinematic equations from the literature. This work has been carried out using the Dymola modeling environment and the Modelica language.

Keywords: omnidirectional wheel, simulation, Modelica, mobile robotics

1. Introduction

Omnidirectional wheels have been invented quite a long time ago [8] and they have a rich history in the literature. They have been used for various tasks and many different embodiments are known [17], [4]. They are constructed so that a vehicle equipped with them can execute true holonomic movements, in other words it can change its direction of movement without changing its orientation. Their great movement capabilities however mean that their mechanical construction is complicated, they also require independent drive and control systems for each wheel. The rolling efficiency of these wheels is worse than that of regular wheels, also they generally do not perform very well on rough surfaces, i.e. they are best suited for indoors applications. Their use ranges from robotic soccer applications, through industrial heavy load transporters [19], and vehicle simulators [1], to educational and entertainment projects like the popular inverted pendulum, but mounted on a ball [12], [2]. Figure 1. shows some of these applications.

1.1. Motivation

To be able to use robotic systems on a professional level, simulation of the individual components and the system as a whole is necessary in the design phase. Modern engineering uses simulation for almost every task imaginable, to cut costs, speed up development and minimize changes late in the product life cycle. Omnidirectional platforms are no exception since they require more complex mechanical design and control, than traditional vehicles.

Figure 1: Examples for the use of omnidirectional wheels[1]

Probably the most important part of a vehicle model is the wheel, since this is the part that makes contact with the ground and transfers forces and torques to move the vehicle. In the last few decades a great number of wheel models of different levels of complexity have been constructed, and are used regularly in automotive and heavy truck simulations. This wealth of knowledge on wheel modeling however has not been applied extensively to other areas of vehicle simulation, such as mobile robotics, although there are a lot of common features between the two.

1.2. Simulation tool - Modelica

Modelica is a free object-oriented modeling language, with a textual definition to describe physical systems in a convenient way, by differential, algebraic and discrete equations. It is supported by the Modelica Association[2]. "It is suited for multi-domain modeling, for example, mechatronic models in robotics, automotive and aerospace applications involving mechanical, electrical, hydraulic and control subsystems, process oriented applications and generation, and distribution of electric power. Modelica is designed such that it can be utilized in a similar way as an engineer builds a real system: First trying to find standard components like motors, pumps and valves from manufacturers' catalogues with appropriate specifications and interfaces and only if there does not exist a particular subsystem, a component model would be newly constructed based on standardized interfaces.

Models in Modelica are mathematically described by differential, algebraic and discrete equations. No particular variable needs to be solved for manually. A Modelica

[1] [12], www.airtrax.com, www.kuka-omnimove.com, [1], [20]
[2] https://modelica.org (Accessed 2012. Feb.).

tool will have enough information to decide that automatically. Modelica is designed such that available, specialized algorithms can be utilized to enable efficient handling of large models having more than hundred thousand equations. Modelica is suited (and used) for hardware-in-the-loop simulations and for embedded control systems." [15] From my point of view the main attractiveness lies in the languages' object oriented nature, which allows a convenient incremental development workflow. Another attractive feature is the model building philosophy of describing the systems by algebraic differential equations, thus approaching the problem from a physics point of view, as opposed to a mathematical one, which – in my experience – is less appealing to an engineer.

1.3. Outline

A configurable omnidirectional wheel model was created that can be adapted to work with most of the popular empirical wheel models used in vehicle simulation today. With the help of this simulation configurable omnidirectional platform models were created and experiments were conducted with the most widely used configurations, the four wheeled Mecanum platform and the three wheeled omnidirectional platform, sometimes referred to as the kiwi drive platform.

The omnidirectional wheel model described in the first part of the article uses a tire model which is available in the Modelica Vehicle Dynamics Library that is included in the academic bundle-version of Dymola 7.4. This model is based on the well-known Rill tire model and handles collision, generates tire forces and torques. My wheel model extends this tire model and creates a roller model which is the basic element of an omnidirectional wheel using two different modeling approaches. In the second part of the paper some simulation results with a four wheeled industrial forklift model and a three wheeled platform are presented. The last part of the paper evaluates the results, gives usage hints and points out future improvement possibilities.

The work was carried out in the Fraunhofer IFF Magdeburg in cooperation with BUTE.

2. Wheel modeling

In this section a short overview is given on the most popular types of omnidirectional wheels and the basic ideas used for modeling them. Later the main concepts of empirical modeling of car tires are highlighted.

2.1. Omnidirectional wheel models

A great number of omnidirectional wheel users fall into the category of industrial companies and users of their product, the most significant being KUKA Robotics[3], and Airtrax[4]. The second group of people who use them fall into the category of students, preparing for robotics competitions such as RoboCup and others, or scientists working on advanced mobility concepts. This second group creates the majority of publications.

[3] http://youbot-store.com/, http://www.kuka-omnimove.com
[4] http://airtrax.com

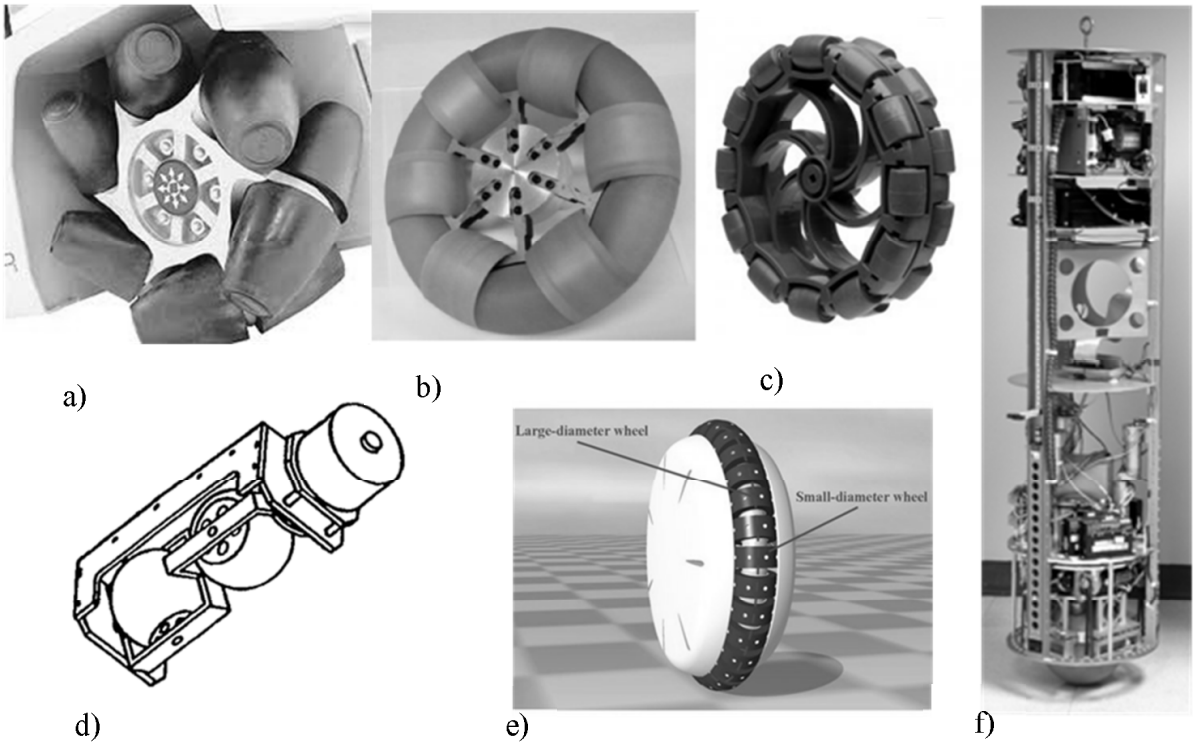


Figure 2 Different omnidirectional wheels and mobility concepts

a) Airtrax Mecanum wheel b) enhanced profile omni-wheel [5] c) multiple row wheel[5] *d) Killough wheel*[6] *e) wheel of a Honda U3-X personal mobility platform f) Ballbot, omnidirectional balancing robot [13]*

Figure 2. shows a cross section of some of the wheel designs and interesting mobility concepts. They were collected to demonstrate relevant concepts trough examples. Subfigures a), b) and c) represent the most popular wheel configurations i.e. passive rollers on the perimeter of a wheel. a) shows the Mecanum wheel used on the Airtrax forklift. It is worth noting that the rollers are shaped in an attempt to attain a round profile and have only a single roller touching the ground at a time. One of the problems associated with omni-wheels is the rough ride associated with changes in wheel radius when changing roller contact. Another important effect is caused by the rigid discontinuities between rollers, they cause slip especially on soft surfaces, such as a carpet [23]. b) and c) are examples of the most widely used solutions to these problems. b) uses different sized rollers, where the larger diameter rollers are shaped so that they can fit the smaller rollers inside, thus virtually eliminating the non-rolling surface on the circumference [5]. This obviously comes at the price of increased complexity. c) is a more common solution, by using multiple regular wheels mounted side by side at an angle, so that "bumpiness" and roller discontinuities can be minimized.

The remaining three subfigures show somewhat different mobility concepts. d) shows the so-called Killough wheel, named after the inventor [17]. In this concept two quasi ball-shaped rollers are mounted in rigid brackets that are connected perpendicular to

[5] http://www.vexrobotics.com
[6] http://www.h33.dk/opfhjul_index.en.html

each other. The rollers are free to roll and the bracket assembly is driven by a motor. These wheels should be mounted and applied just like the omni-wheels above. (More on Kinematic constraints can be read in [9]) Smooth ride and single roller contact is ensured by the shape of the rollers, the contact point however moves significantly when roller contact changes.

Subfigure e) shows the wheel of Honda U3-X[7] personal mobility platform. A seat is mounted on top of the wheel and the vehicle balances and drives on this single wheel in an omnidirectional fashion. This is achieved by powered rollers in addition to the main drive that turns the entire wheel.

f) shows an omnidirectional platform that clearly eliminates any rolling imperfections by using a ball to ride on. It is called Ballbot and it was designed to work in areas used by people [13]. It is high enough to make eye contact yet it has a small footprint, that together with omnidirectional maneuverability enables it to get around in cluttered indoor environments.

My model can be used to describe the type of wheels represented by the first three subfigures. A common characteristic is that they have a relatively small width relative to their diameter and they are designed with an attempt to ensure smooth ride. In the following let us take a look at how this type of wheels has been modeled by other researchers.

Williams et al. [23] developed a wheel model motivated by the RoboCup competition. They used a small three wheeled platform with 0° rollers (Figure 3. a). The wheels they used had a single row of rollers, without any provisions to smoothen roller discontinuities, this was reflected in their results, the wheels demonstrated a strong angle dependent friction characteristics directly related to the non-rolling part touching the carpet they used for testing. It is also important to note that they found that the friction coefficient in the driven and in the free rolling direction was comparable – 3/1 and 5/3 respectively for paper and carpet – showing that the quality of the omni-wheel greatly effects the behavior of the mobile platform.

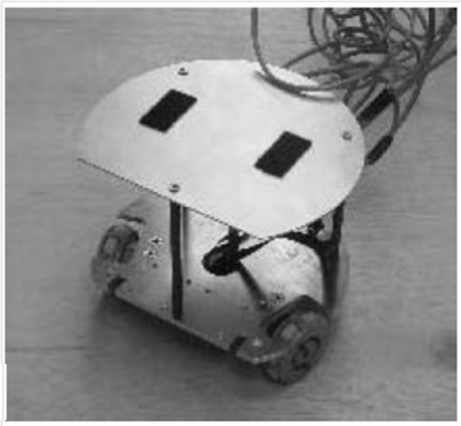
a)  b) 

Figure 3 a) Omnidirectional RoboCup player by Williams et al. [23], b) youBot by KUKA Robotics, flexible arm on a mobile base

[7] http://www.hondanews.com

Dresscher et al. [6] developed a modular model for youBot (Figure 3. b) using an energy based method and its bond graph representation. Their goal was to model this rather complicated platform in a modular, reusable fashion, this also includes modeling the Mecanum wheels. To make the model simpler they neglected the dynamic behavior of the wheels and derived a kinematical model from the geometry. They neglected friction in the roller bearings and generally neglected force in the free rolling direction. They defined a transformation between drive axis movement and wheel movement in the direction parallel to the roller axis. Floor contact was modeled with a resistance and a stiffness parameter. The authors had no opportunity to validate their model on the real platform.

Tobolár et al. [21] created an object oriented library for Mecanum wheels in Modelica, unfortunately their article is very short and non-informative. As the author explained this is due to an NDA with KUKA Robotics.

Studying the literature the conclusion can be drawn, that in many cases omnidirectional platforms are modeled as a whole, assuming symmetrical load distribution, without having separate wheel models. However when wheels are modeled, dynamic effects are often neglected and the results are purely kinematical. This is probably justified when the platform has very low, known weight, for example a RoboCup player. Another common modeling approach is that wheel forces are assumed to be generated parallel to the direction of the rollers and forces perpendicular to the roller axis are assumed to be zero. An exception is the paper mentioned before [23], where the authors had to calculate with substantial forces in the free rolling direction, however in my opinion this was due to the disadvantageous characteristics of the omni-wheel they used.

To be able to apply a wheel model that accommodates a broad range of robotic platforms including heavy machines, with uneven load distribution and various wheel designs with different roller materials, a model is needed that is easy to parameterize, includes simple dynamics, handles sliding and last but not least of all, well suited for simulation. For this a choice was made to build on the well proven results from the domain of regular tire modeling, while applying some of the techniques used in the literature cited above.

2.2. Regular tire models in simulation

Tire modeling in general has been an active area of research for a long time, because the behavior of a tire is a complex phenomenon, and the results can be used in countless applications, making it both a challenging and lucrative area of research. The main purpose of a tire, besides providing a smooth ride for the passengers is to transmit forces and torques in three mutually perpendicular directions to create vehicle movement and directional control. To achieve this, a tire model has to handle collision, calculate the contact patch with the ground and obstacles, and it has to generate the forces and torques that arise. Most of these calculations are nonlinear because of the characteristics of the tire material [3].

Instead of creating a model from scratch, already existing models in Modelica were used as a basis, thus not having to recreate well proven components.

The simplest models regard the tire as a rigid disk, with unchangeable radius and linear dynamic properties, the most complicated ones use finite element simulation, fine tuned to a certain rubber compound and carcass. For a tire model to be useful, a compromise between complexity and accuracy has to be found depending on the application at hand. A very good example of incremental model building in Modelica is given by [24].

Except for simple targeted experiments the model cannot be restricted to a certain driving situation. Most of the relevant cases have to be considered, such as driving with nonzero camber and sideslip angles. Another important aspect is the adaptability of tire model characteristic parameters to real world tire behavior. Most experimental tire models such as the well-known Magic formula by H. Pacejka, or TMeasy by G. Rill [7], [16], [18] uses polynomial approximation of real measured data curves. These polynomials and the conditions of switching between them describe the tire characteristics, for given external parameters. An example according to the Rill model is given in Figure 4. The curve is valid for constant coefficient of friction μ and vertical load F_Z.

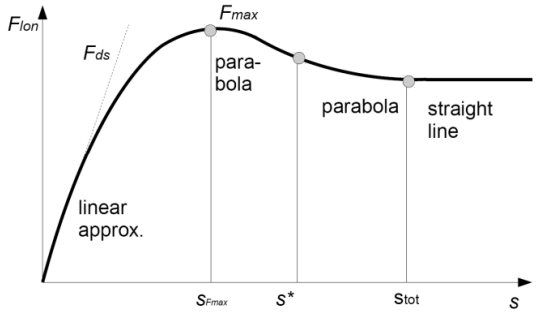


Figure 4: Steady state longitudinal force vs. wheel slip approximation according to [7] for a given μ and vertical load

Slip is usually defined in relation to the difference between a wheel center's velocity and its circumferential velocity. It is often confused with sliding, but it is important to see that slip occurs with a perfectly gripping tire as well. It can be thought of as the driven axis twisting the elastic tire material around itself.

2.2.1. The Rill tire model

A very appealing aspect of the Rill model is that its parametrization is intuitively simple, containing only a few parameters, with physical meaning, in contrast to for example the magic formula, which uses a lot of curve approximation parameters with no direct physical meaning. The parameter values matching measurement data are available from certain manufacturers for some of their products. Nevertheless one would need to match experimental data again when creating a new model for an unknown tire, thus restricting the usage of this kind of modeling to users well equipped with tools for tire identification.

The Rill model is a semi-empirical tire model with first order dynamics. Wheel parameters for steady state force generation are approximated quadratically from two load tables: one similar to Table 1 - for a nominal F_Z - and another one for $2F_Z$. Table 1

shows some typical values used in the Modelica libraries and [18], for reference. The indices x and y stand for longitudinal and lateral values, respectively.

Table 1: Load table for Rill model and typical values

Name	Description	Typ. value
F_{znom1}	Nominal normal force	3000N
F_{ds_x1}	Slope at $s_x = 0$	50000N
s_{max_x1}	Slip of maximum tire force	0.15
F_{max_x1}	Maximal tire force	3000N
s_{slide_x1}	Slip where sliding begins	0.4
F_{slide_x1}	Force where sliding begins	2800N
F_{ds_y1}	Slope at $s_y = 0$	40000N
s_{max_y1}	Slip of maximum tire force	0.21
F_{max_y1}	Maximal tire force	2750N
s_{slide_y1}	Slip where sliding begins	0.6
F_{slide_y1}	Force where sliding begins	2500N

The other load table contains values for $2F_{znom1} = F_{znom2}$ *and the quadratic interpolation for* $F_z < 2F_{znom1}$ *for any value is demonstrated trough the example of* F_{ds_x}:

$$F_{ds\_x} = \left[\left(\frac{F_{dsx_2}F_{znom1}}{F_{znom2}} - \frac{F_{dsx_1}F_{znom2}}{F_{znom1}}\right)\left(\frac{F_{znom1} - F_{znom}}{F_{znom1} - F_{znom2}}\right) + \frac{F_{dsx1}F_{znom}}{F_{znom1}}\right]\frac{F_{znom}}{F_{znom1}} \quad (3)$$

The interpolation curve for given parameter values is shown on Figure 5.

Besides the tire load parameters some dynamic parameters also have to be set. The model uses first order dynamics counting with a dynamic force according to the following - in direction x:

$$F_{dyn\_x} = c_x e_x + d_x \dot{e}_x \quad (2)$$

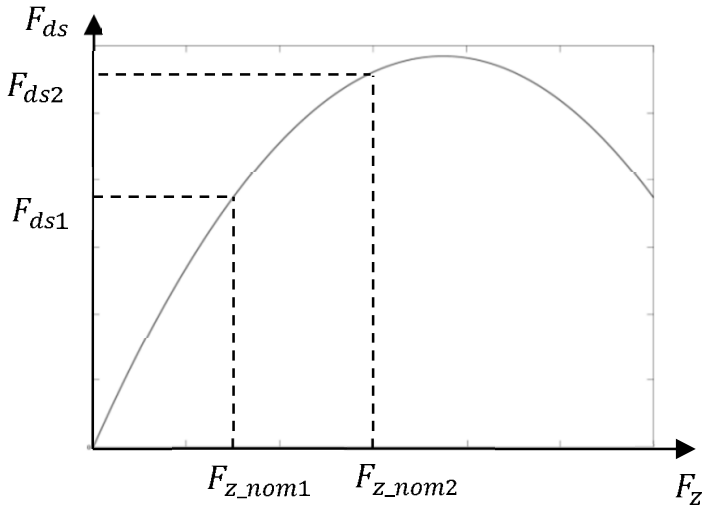


Figure 5: Force curve gradient quadratic interpolation example, F_{ds_x} *vs. vertical load*

Meaning of the parameters is explained in Figure 6 according to [18], where e_x is the longitudinal tire deformation and c_x and d_x are the lateral stiffness and damping, respectively. Typical values for c_x and d_x are 100000 N/m and 1500 Ns/m. In the z and y directions the behavior is similar, however the constants might differ.

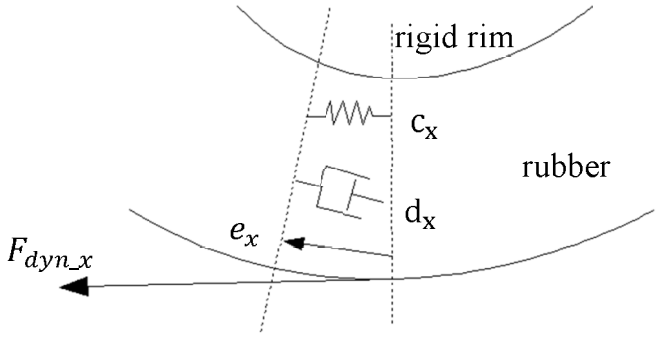


Figure 6: First order tire dynamics

2.3. Omnidirectional models based on regular tires

Two distinct approaches were used to adapt regular models to omnidirectional operation. These are documented in the following sections. Both models can be adapted to use various tire models – not just the Rill model – as a basis.

2.3.1. Using individual rollers

The most straightforward method to create a usable Mecanum wheel model:

- take any tire model from the library to create a roller from the base model
- set estimated (or measured) wheel parameters and geometry
- set a certain number of rollers and arrange them according to the given wheel geometry
- allow the rollers to spin freely along their main axis and connect them to a main axle, that can be driven by an appropriate angular velocity or torque

This is illustrated on Figure 7, with a Mecanum wheel, where "frame_a" is the fixed wheel hub, the wheel is driven by the "speed" variable (angular velocity) "rollerRot" creates the 45° rotation and "spoke" defines the position of each roller wheel around the perimeter. The number of rollers is configurable trough a user-defined parameter.

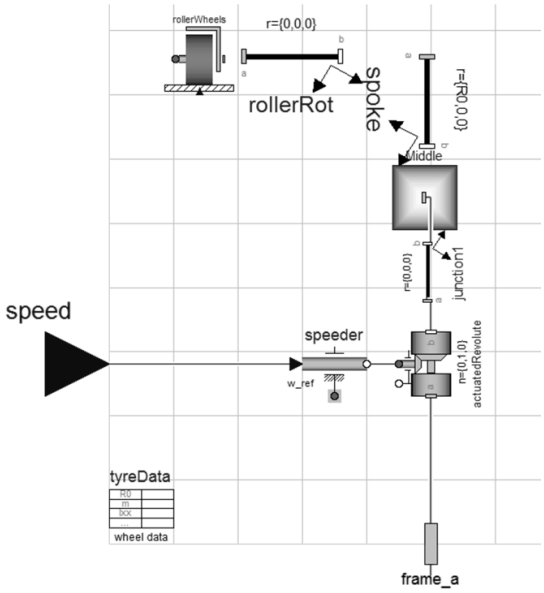
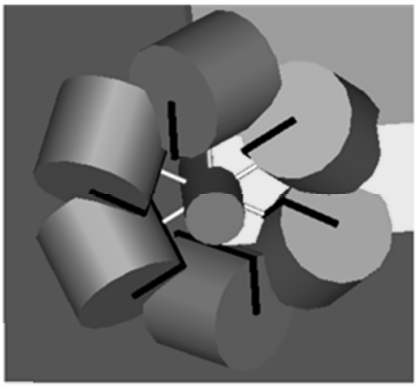


Figure 7: Mecanum wheel model of an individual roller in Modelica

An example animation result for six rollers can be seen on Figure 8 a). The spokes – basically the vectors pointing at rollers' centers – are computed according to Figure 8 b). Each wheel's local x-direction points forwards while the z-axis points upwards. Together with the y axis these create a right handed coordinate system.

a)



b)

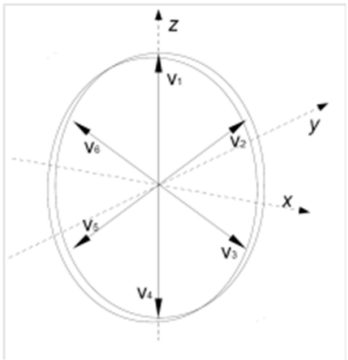


Figure 8: Animation a) and virtual spokes b) of a Mecanum wheel

As the wheel lies in the x-z plane, the spoke vectors – starting at the wheel's origin – are obtained the following way:

$$v_i = R_0 \left\{ \sin\frac{2\pi i}{n}, 0, \cos\frac{2\pi i}{n} \right\} \tag{3}$$

where i goes from 1 to n which is the number of rollers and R_0 is the spoke length.

This approach is straightforward, since it copies wheel mechanics, and it does work fairly well in simulation:

- Straightforward implementation.
- Very easy to switch between different tire models.
- Implicitly handles roller inertia, and rolling resistance.
- The model demonstrates "bumpiness" when it rolls from a roller to another, real mecanum wheels cause a bit of unevenness when rolling too.
- If simulation time is not an issue, adding more rollers and/or a better contact geometry model could make it more realistic.

It also suffers from several disadvantages.

- Far from suitable for real time simulation. Complicated model - for a typical four wheeled six roller vehicle, collision detection and force calculation has to be carried out for 24 rollers.
- Relies on boundaries of original wheel model. The individual rollers operate at extreme situations: up to 90° sideslip and camber angles. The tire model can handle this, but it was not designed for it.
- Crude contact model, most Mecanum wheel rollers are conical. In order to make them ride smoother, they have a varying cross section and rounded edges instead of being regular cylinders. A better geometry model would add complexity (see first point).

Naturally all these problems can be solved, however at a price of violating the principle of the original goal of creating a simple yet realistic wheel model, using available components. This approach would need a new contact model and a new roller design, from the start.

2.3.2. Single roller model

To overcome some of the disadvantages of the model presented above another one was created, based on a different approach. The main idea is to alter the force generation method of a single tire to behave like an omnidirectional wheel, applying some of the basic ideas summarized in section 2.1.

For a real Mecanum wheel the number of rollers touching the ground varies between one and two, creating an angle dependent effect on the wheel forces. However, in this model we assume that the force generation is continuous along the perimeter of the wheel much like an extrapolation of the ideal case when the center of only a single roller touches the ground. This is reasonable as the rollers are usually shaped in an attempt to achieve this effect. (see Figure 2)

To describe the modifications let us introduce the notation system used in the Rill model and the Modelica model for a regular wheel. They use the C (carrier) and W (wheel) coordinate systems according to the TYDEX [22] notations. "The C-axis system is fixed to the wheel carrier with the longitudinal xc-axis parallel to the road and in the wheel plane (xc-zc-plane). The origin of the C-axis system is the wheel center.

The origin of the W-axis system is the road contact-point defined by the intersection of the wheel plane, the plane through the wheel carrier, and the road tangent plane".[8]

The unit vectors Ce_x, Ce_y, Ce_z and We_x, We_y, We_z, point in the direction of the C and W system axes. To accommodate the omnidirectional wheel, we define a We_{fw} unit vector in the direction of the rollers' axis that is the direction it can exert force (see Figure 9)

$$We_{fw} = We_x \cdot Rot_\delta \tag{4}$$

where Rot_δ is a 3x3 rotation matrix of δ.

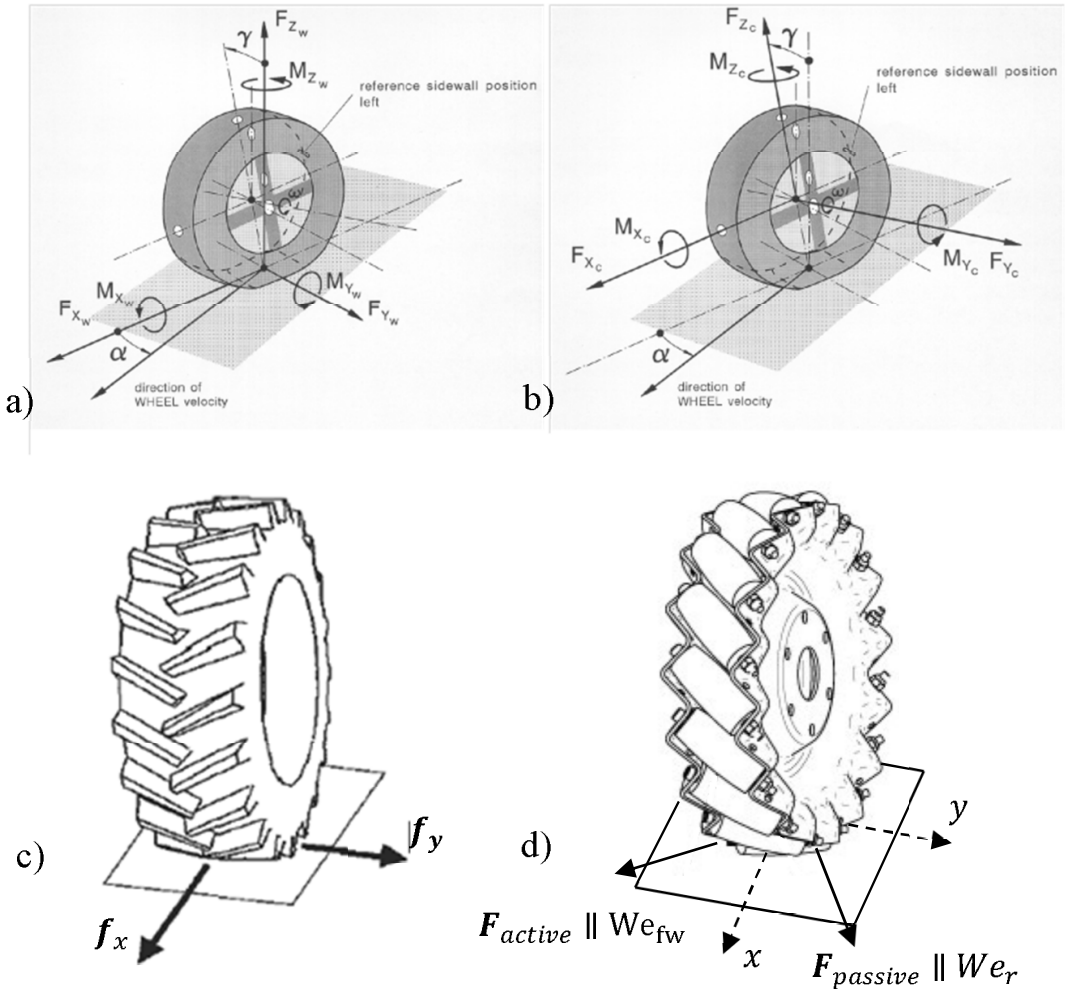


Figure 9: Definition of roller vectors a)Tydex C-axis system b) Tydex W-axis system c) Forces in the ground plane, regular wheel d) Forces in the ground plane omni-wheel

For the sake of simplicity we assume that the wheel does not exert any force to the direction of We_r - that is the free rolling direction perpendicular to We_{fw} . A more detailed model could include roller inertia and/or rolling resistance. Having made this assumption, we can make a further step by defining slip for the Mecanum wheel. In most wheel models forces are generated as a function of slip, so this is an important

[8] http://ti.mb.fh-osnabrueck.de/adamshelp/ (accessed May 2012)

aspect. Slip is defined separately for the x and y directions. Since our idealized roller only generates force in the direction of its spin axis (We_{fw}) we shall only calculate slip in this direction. [18] defines slip as "total slip":

$$s = \frac{v}{R|\Omega| + v_{num}} \quad \text{where} \quad v = \sqrt{v_x^2 + v_y^2} \tag{5}$$

R is the wheel radius, Ω is the angular velocity, and V_{num} is a small number inserted for numerical reasons. In our model we modify v in the slip equation:

$$v_{mecanum} = \sqrt{v_{fw}^2} \tag{6}$$

where v_{fw} is the projection of the velocity of the center of the wheel in the We_{fw} direction.

After redefining the slip equation, all we need to do is equate static and dynamic force equations with zero in the y direction and calculate force in the x direction according to the Rill model using the modified slip equation. The direction of this force has to be set to the direction of We_{fw}. At this point the effects of camber, rolling resistance and bore torque were not investigated.

3. Usage

The model described in this paper was used to simulate an industrial robotic transport vehicle, a forklift with a Mecanum wheel and a three wheeled robot with $\delta = 0^o$ omni-wheels. To use a Mecanum wheel, the most popular configuration is to put four of them on a platform in a way so that the rollers on the ground form a rhomboid, for the other platform a symmetric configuration is preferable Figure 10. (Other configurations, using more wheels and asymmetry are possible, for a general kinematics discussion see [9])

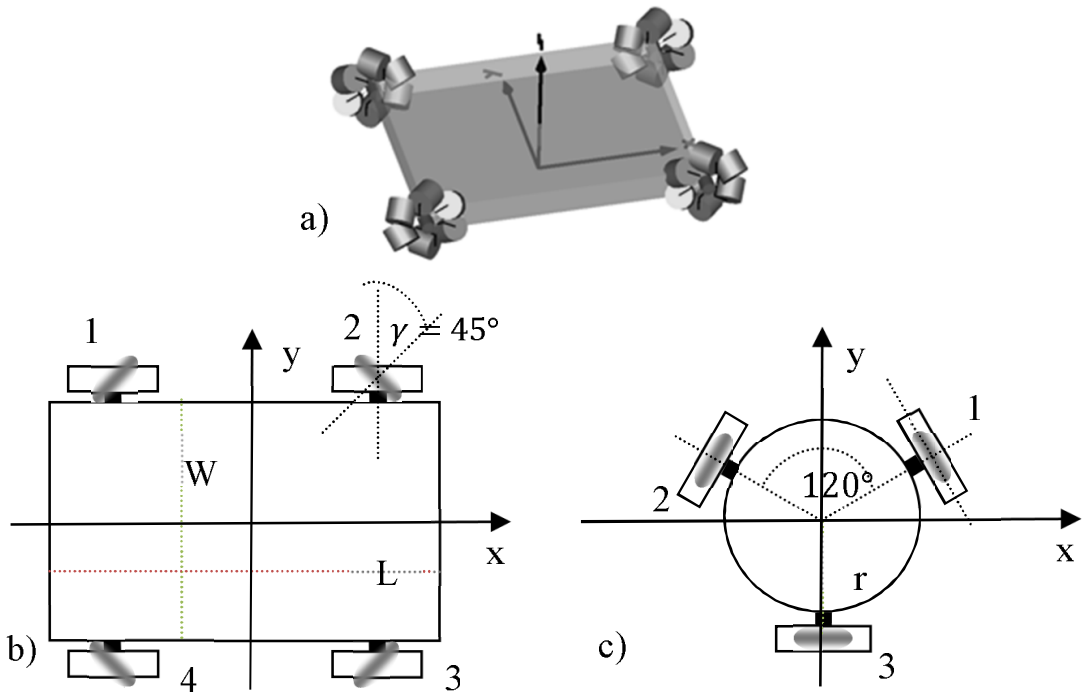


Figure 10 Illustration of wheel setup for two popular omnidirectional platforms

To obtain the correct wheel setup the front left and the rear right wheels have to be mirrored meaning that for them

$$\mathrm{We_{fw}} = \mathrm{We_x} \cdot \mathrm{Rot_{-45°}} \tag{7}$$

Figure 10 a) depicts a model with the separate rollers model (section 2.3.1) for better visibility of wheel orientation, naturally the same configuration is used in the single-roller model (section 2.3.2) with - δ rotation. The virtual experiments were carried out on a flat surface, therefore to keep it as simple as possible, no suspension was simulated. The platform body was represented by a mass with inertia, and the load by a point mass.

Wheel parameters can be set intuitively or by making simple measurements. At the time of writing the author had no means to identify experimental parameters, so only intuitively set values were taken. Basically a stiff wheel with small carcass was used to model the lack of an inflated tire body.

The general kinematic equations for controlling the platform are well known from the literature [14], [9] Mecanum platform:

$$\begin{pmatrix}\omega_1\\ \omega_2\\ \omega_3\\ \omega_4\end{pmatrix} = \frac{1}{\mathrm{R}}\begin{pmatrix}1 & 1 & -(\mathrm{L}+\mathrm{W})\\ 1 & -1 & -(\mathrm{L}+\mathrm{W})\\ 1 & 1 & (\mathrm{L}+\mathrm{W})\\ 1 & -1 & (\mathrm{L}+\mathrm{W})\end{pmatrix}\begin{pmatrix}v_x\\ v_y\\ \Omega\end{pmatrix} \tag{8}$$

Where the ω-s designate angular velocities for each wheel numbered according to Figure 10, R is the wheel radius, L and W are length and width of the platform, v_x,v_y are platform velocities in local coordinates and Ω is the platform angular velocity.

Three wheel platform:

$$\begin{pmatrix}\omega_1\\ \omega_2\\ \omega_3\end{pmatrix}=\frac{1}{R}\begin{pmatrix}0.5 & -\frac{\sqrt{3}}{2} & -r\\ 0.5 & \frac{\sqrt{3}}{2} & -r\\ -1 & 0 & -r\end{pmatrix}\begin{pmatrix}v_{c_x}\\ v_{c_y}\\ \Omega\end{pmatrix} \tag{9}$$

By using these well known equations we were able to verify whether our wheels move the platform as expected.

Figure 11 a) shows an experiment with single-roller Mecanum wheels. The arrows at the wheels represent the driving force generated at the contact point.

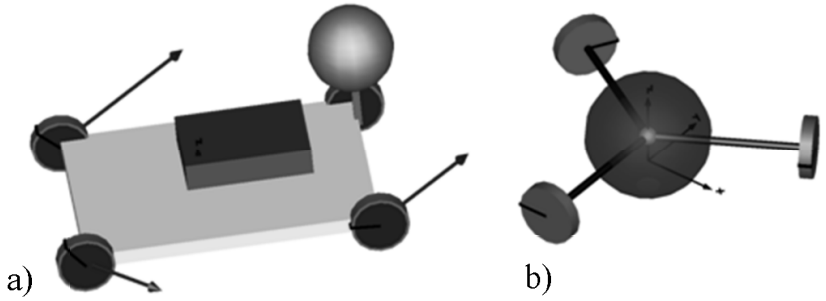


Figure 11 a) Forklift model with single roller Mecanum wheels b) Three wheeled platform with 0° omni-wheels

Figure 12 demonstrates the maneuverability of the platforms and the usability of the wheel model. Captured moments of a turn-while-translate movement can be seen (sampling not equidistant, for better visibility).

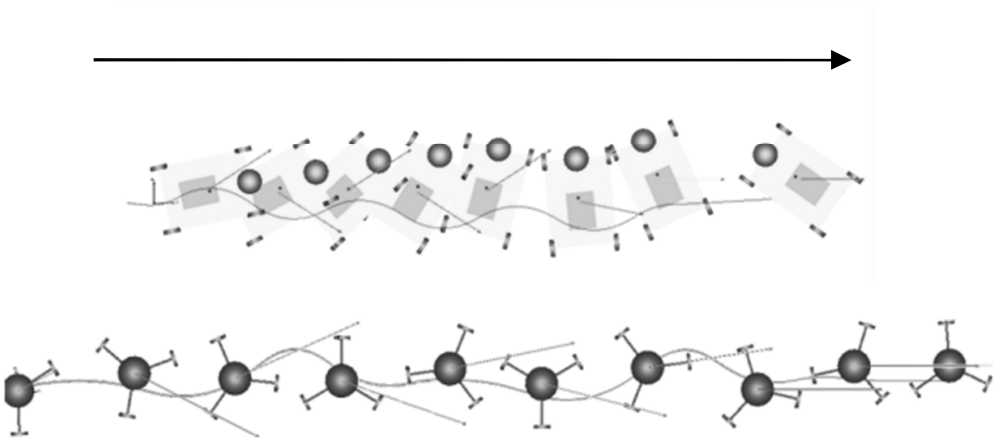


Figure 12 Turn-while-translate maneuver with zigzag movement

The platform continuously changes its orientation while translating in x and y directions simultaneously. The blue curve represents the trajectory of a chosen chassis point. The instantaneous speed vector is displayed as a green arrow.

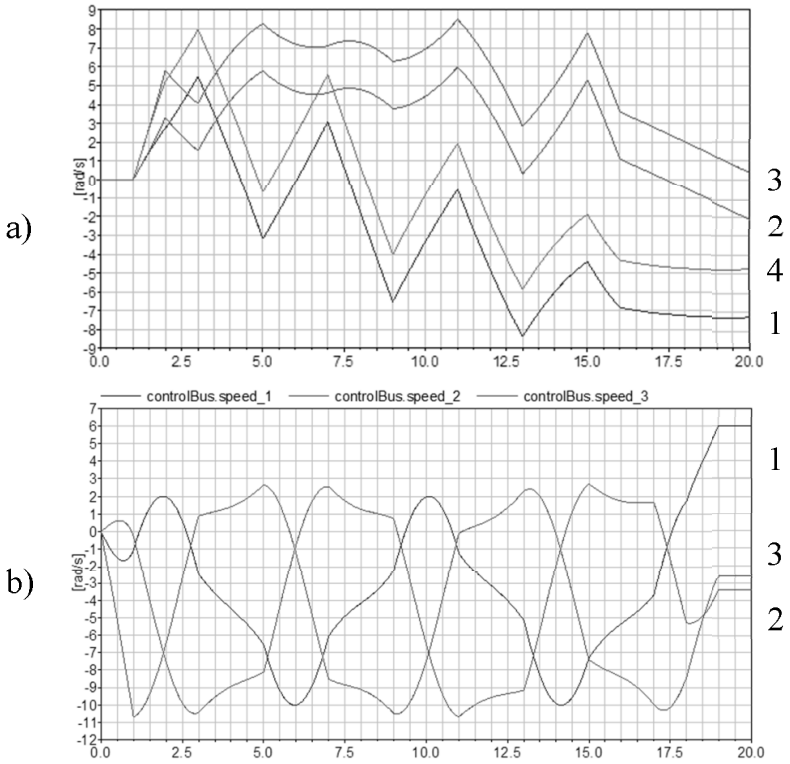


Figure 13 Angular velocity commands vs. time for a) Mecanum b) three wheel platforms while performing a turn-while-translate maneuver with zigzag

Figure 13 shows the angular velocity control signals versus time, exported from Dymola for each wheel. The angular velocities are constantly changing, to create the prescribed motion pattern.

These robot models can be used for trajectory generation and controller design tasks, as well as dynamic behavior tests for safety applications. A slip-based center of gravity estimation method is described in [11], and a brake assistant for omnidirectional wheels is published in [10]. For these works the single roller omnidirectional wheel model was used.

4. Scope - future prospects

The model follows the structure used by the Modelica library so integration into existing designs is simple. At this point the model has only been used for experiments on flat ground and zero camber angle. As real omnidirectional wheels are usually used under these conditions it is a reasonable simplification. However, modeling the effect of small obstacles and ground inclinations could be an interesting research topic. In my second – single roller – model the physical effects due to multiple individual rollers, were totally neglected. In order to make the wheel model more accurate, further dynamic effects such as roller inertia, rolling resistance could be incorporated. A possibly more important factor is the consideration of roller discontinuities. When two adjacent rollers come into contact with the ground, they might cause small fluctuations in effective wheel radius, contact location and also change the slip characteristics when multiple contact points occur. This effect could be incorporated by multiplying existing force characteristics with an angle dependent function, related to this effect. The

omnidirectional wheels that can be reliably modeled by my method are those that have free rolling rollers on their circumference, shaped in an attempt to achieve smooth ride and small contact point fluctuations. Further plans include verifying and calibrating the model to a real mobile platform.

5. Conclusion

In this paper a method for modifying an existing, widely used tire model was presented. By this modification the model is able to describe the force generation of a class of omnidirectional wheels. The wheel model was implemented in Modelica in two different embodiments, for two different platforms. The usability of the models was demonstrated by applying them on an industrial forklift model. The model was created with simplicity and ease of use in mind, so some effects of smaller importance are not modeled. Some hints for the future extension of the wheel model were given at the end of this article.

The author would like to express his appreciation to Dr. Tamás Juhász and people at the IFF Fraunhofer Magdeburg who helped him with Modelica and received him at their lab. He would also like to thank Dr. László Vajta for valuable suggestions and inspiration for this work.

References

[1] R. Ahmad, P. Toonders, M.J.D. Hayes, and R.G. Langlois. *Atlas mecanum wheel jacobian empirical validation.* In *CSME International Congress*, Winnipeg, MA, Canada, 2012.

[2] Magnus Jonason Bjärenstam and Michael Lennartsson. Master's thesis, Lund University, Department of Automatic Control, 2012.

[3] Raymond M. Brach and R. Matthew Brach. *Tire models for vehicle dynamic simulation and accident reconstruction.* Technical report, Brach Engineering, 2009. SAE Technical Paper.

[4] Jochen Brunhorn, Oliver Tenchio, and Raúl Rojas. Robocup 2006: *Robot soccer world cup x. chapter A Novel Omnidirectional Wheel Based on Reuleaux-Triangles*, pages 516–522. Springer-Verlag, Berlin, Heidelberg, 2007.

[5] Kyung-Seok Byun and Jae-Bok Song. Design and construction of continuous alternate wheels for an omnidirectional mobile robot, *Journal of Robotic Systems*, 20(9):569–579, 2003.

[6] Douwe Dresscher, Yury Brodskiy, Peter Breedveld, Jan Broenink, and Stefano Stramigioli. *Modeling of the youbot in a serial link structure using twists and wrenches in a bond graph.* In *Proceedings of SIMPAR 2010 Workshops*, pages 385–400, Germany, November 2010. SIMPAR.

[7] W. Hirschberg, G. Rill, and H. Weinfurter. *Tire model TMeasy. VEHICLE SYSTEM DYNAMICS*, 45(S):101–119, 2007.

[8] Bengt Erland Ilon. *Wheels for a course stable selfpropelling vehicle movable any desired direction on the ground or some other base*, 1975. US Patent No. 3876255.

[9] Giovanni Indiveri. *Swedish wheeled omnidirectional mobile robots: Kinematics analysis and control. IEEE Transactions on Robotics*, 25:164 – 171, 2009.

[10] Viktor Kálmán and László Vajta. *Designing and tuning a brake assistant for omnidirectional wheels, Periodica Polytechnica* 56:(4), 2012.

[11] Viktor Kálmán and László Vajta. *Slip based center of gravity estimation for transport robots.* In *Factory Automation*, pages 50–55, Veszprém, Hungary, May 21-22. 2012. University of Pannonia.

[12] Masaaki Kumaga and Takaya Ochiai. *Development of a robot balanced on a ball - application of passive motion to transport.* In *Robotics and Automation, 2009. ICRA '09. IEEE International Conference on*, pages 4106 –4111, may 2009.

[13] T.B. Lauwers, G.A. Kantor, and R.L. Hollis. *A dynamically stable single-wheeled mobile robot with inverse mouse-ball drive.* In *Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on*, pages 2884 – 2889, may 2006.

[14] P. Muir and C. Neuman. *Kinematic modeling for feedback control of an omnidirectional wheeled mobile robot.* In *Robotics and Automation. Proceedings. 1987 IEEE International Conference on*, volume 4, pages 1772 – 1778, mar 1987.

[15] Martin Otter and Hilding Elmqvist. *Modelica - Language, Libraries, Tools, Workshop and EU-Project RealSim.* German Aerospace Center, Dynasim AB, June 2001.

[16] Hans B. Pacejka. *Tyre and vehicle dynamics.* Butterworth-Heinemann, 2002.

[17] F.G. Pin and S.M. Killough. *A new family of omnidirectional and holonomic wheeled platforms for mobile robots. Robotics and Automation, IEEE Transactions on*, 10(4):480 –489, aug 1994.

[18] Prof. Dr.-Ing. Georg Rill. Vieweg+Teubner Verlag, 1994.

[19] Daniel Ruf and Jakub Tobolár. Omnidirektionale fahrzeuge für schwerlasttransport in produktion und logistik. *Logistik und Verkehr in Bayern*, 12:34–35, 2011.

[20] D. Stonier, Se-Hyoung Cho, Sung-Lok Choi, N.S. Kuppuswamy, and Jong-Hwan Kim. *Nonlinear slip dynamics for an omniwheel mobile robot platform.* In *Robotics and Automation, 2007 IEEE International Conference on*, pages 2367 – 2372, april 2007.

[21] J. Tobolár, F. Herrmann, and T. Bünte. *Object-oriented modelling and control of vehicles with omni-directional wheels.* In *Computational mechanics 25th conference with international participation*, Hrad Nectiny, Czech Republic, November 9-11 2009.

[22] H.-J. Unrau and J. Zamow. *TYDEX-Format, Description and Ref. ManualTYDEX-Format, Description and Ref. Manual.* Initiated by the TYDEX Workshop, release 1.3 edition, Sept. 1997.

[23] II Williams, R.L., B.E. Carter, P. Gallina, and G. Rosati. *Dynamic model with slip for wheeled omnidirectional robots. Robotics and Automation, IEEE Transactions on*, 18(3):285 –293, jun 2002.

[24] Dirk Zimmer and Martin Otter. *Real-time models for wheels and tyres in an object-oriented modelling framework. Vehicle System Dynamics*, 48(2):189–216, 2010.

The Significance of Developing A New Black Spot Safety Management Approach for The Local Road Traffic Nature of Ho Chi Minh City

H. Nguyen, P. Taneerananon, C. Koren, P. Iamtrakul

e-mail: huycongtrinh@yahoo.com, breathislife@yahoo.com, koren@sze.hu, apawinee@hotmail.com

Abstract: Road networks always have inherent levels of risk. Whenever a person is driving on a roadway, the risk of being involved in a collision exists. This situation became worse with a poor safety engineering road network as in Ho Chi Minh City (HCMC). In HCMC, every year, there are more than one thousand deaths and approximately ten thousand road users were injured due to road traffic accidents. Until now, in Vietnam, there has been almost no road safety method used in management of road network safety. This paper was intended to propose an approach of black spot safety management suitable for local road traffic nature of HCMC. The approach developed is based on the *Network Safety Management (BASt & Sétra, 2005)*. The methodology employed in the approach focuses on the traffic volume and the severity of accidents within the road network and the evaluation of the accidents on the basis of accident cost rates. The comparison of actual accident cost with a hypothetical estimated base accident cost provides information on safety potential (SAPO) of spots. The SAPO is the most important parameter to identify black spots on which safety improvement measures are expected to have the greatest effect. Furthermore, some concepts and definitions related to the approach are also described in detail.

Keywords: black spot management, black spot identification, black spot analysis

1. Introduction

Black spot safety management (BSM) approach was proven a very efficient method in improving road traffic safety when many accidents of the same type happened at more or less the same site in the road network. However, the traditional BSM approaches usually relied on recorded number of accidents, critical values, and the sliding window method to identify black spots. As a result, the set of hazardous road locations identified will contain both true and false positives. This led to the fact that the analysis may include a number of sites which are not really necessary to detect, and a number of sites with false negatives may go undetected. This research introduces a new black spot management approach which is based on the SAPO in identifying and ranking black spots in order to maximize the effectiveness of safety improvement measures.

2. Background

HCMC is the biggest financial and economic hub of Vietnam, linking the Southern areas with the other parts of the country as well as with foreign countries. The demand for the city's transport is rapidly growing. However, the transport infrastructure, especially the road sector, remains very poor and fails to keep in pace with the development growth.

Source: JICA, 2006

Figure 1. Inter-regional road network and the main road network in HCMC and surrounding areas

The current road networks in HCMC and surrounding regions are as in Figure 1. All of the national roads either starting or ending in HCMC, connect this city with the surrounding regions, and connect these regions with one another. In addition, the provincial road networks are of poor quality and ineffectively connect the centres of the districts with national roads.

3. General information on road traffic safety in HCMC

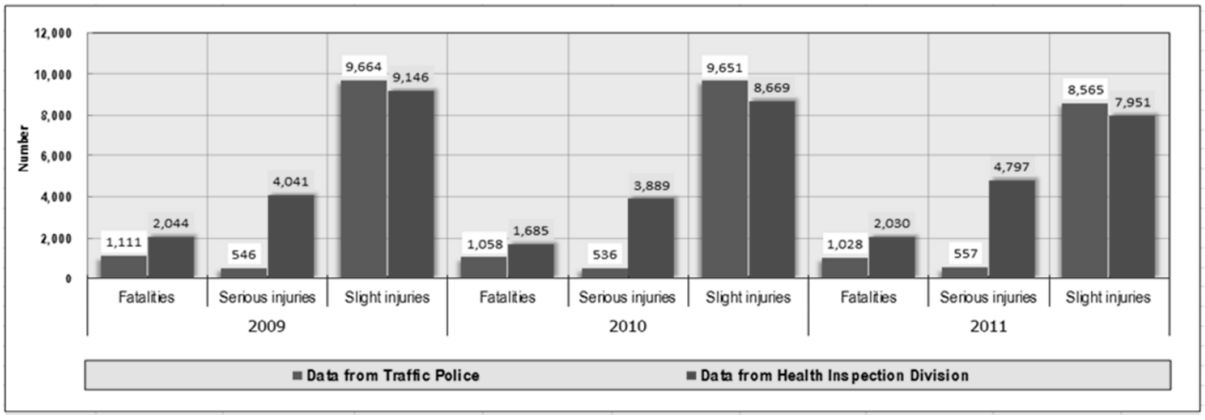


Figure 2. Comparison between two accident data sources in HCMC (2009-2011)

Figure 2 shows the difference in quantity of road traffic accidents reported by the two sources. This means that a considerable number of accident cases are under-reported.

Actually, the number of serious injury and fatality cases recorded by HCMC Health Inspection Division is much higher than that by HCMC Traffic Police.

3.1. Accidents according to road types

Owing to heavy traffic with strong conflict between vehicles of different sizes and speeds, the percentage of road accidents happening on urban roads is relatively higher than that of accidents happening on sub-urban roads and national roads. The distribution of road accidents according to road types in the city in the period from 2009 to 2011 is illustrated as in Figure 3.

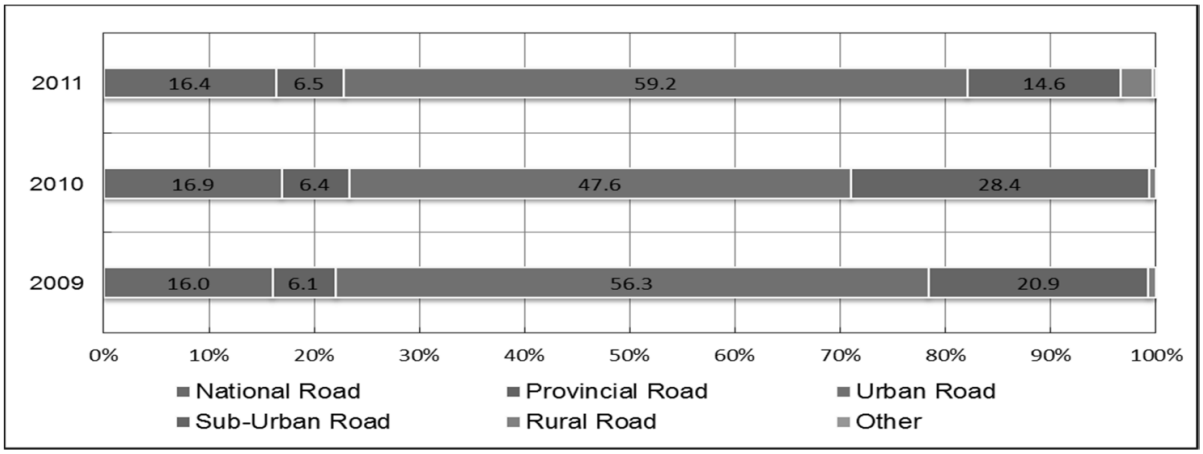


Figure 3. Road traffic accidents according the type of roads (2009-2011)

3.2. Accidents according to vehicle types

The prominent feature of vehicle types in HCMC is that motorcycles account for 90% of the total number of vehicles. As a result, this type of vehicles takes up 90% of the road traffic capacity of the city. 70% of the road accidents are caused by motorcycles, 15% by taxis and 5% by trucks. The remaining 10% are caused by other types of vehicles such as trailers, buses, coaches, and so on. The distribution of road accidents according to vehicle types in the city in the period from 2008 to 2010 is illustrated in Figure 4.

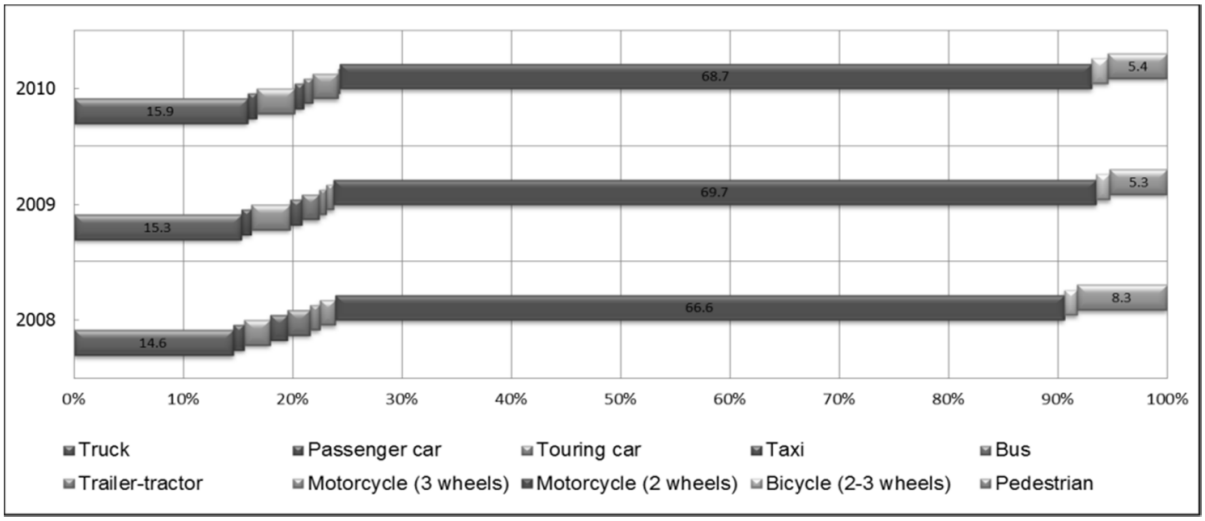


Figure 4. Road accident according to type of vehicles (2008-2010)

3.3. Accidents according to crash types

Most of the road accidents are crashes between motorcycles, accounting for a percentage of approximately 60%. Next comes, the case of between motorcycles and cars, accounting for a percentage of 18%. The remaining 22% are the cases of car-car accidents, motorcycle-pedestrian accidents, and so on. The distribution of road accidents according to crash types in the city in the 2010 is illustrated in Figure 5.

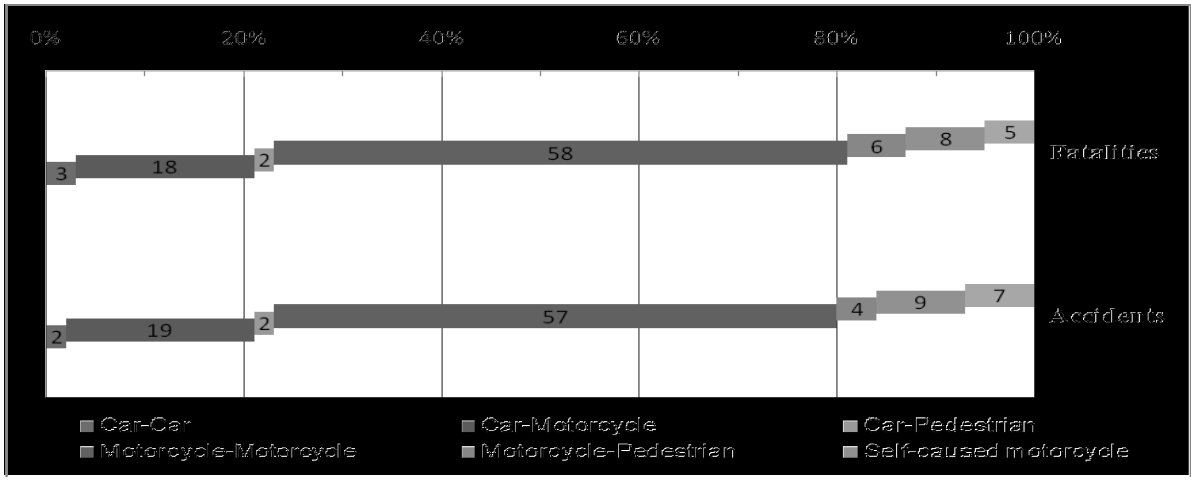


Figure 5. Accidents according to crash types (2010)

3.4. Accidents according to the age of road users

According to the statistics for road accidents in 2010, most of the accidents are caused with the involvement of road users of the age between 19 and 40. The distribution of road accidents according to the age of road users in the 2010 is illustrated as in the figure below. The Figure 6 shows that most of cases of death and serious injury in road accidents in HCMC happen to people who are still at the age of labour. These people are often the breadwinner in their families and also the main labour force the society.

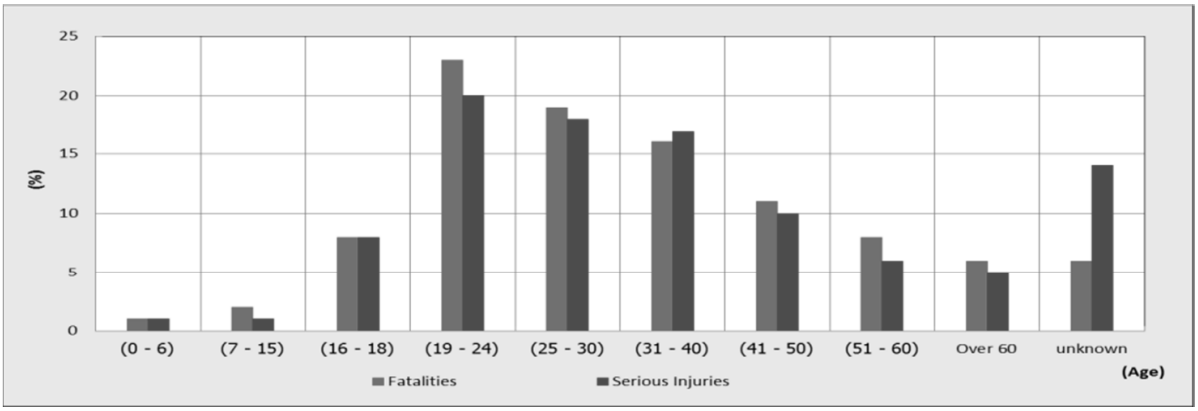


Figure 6. Accidents according to age of road users in the year 2010

4. The significance of developing a new BSM approach

There have been so many accidents occurring on the road network of poor safety and engineering. However, there has not been any road safety methods effectively used in management of road network safety in HCMC. Therefore, a new BSM approach should

be proposed to improve the road traffic safety in this city. The following sections will describe the approach in detail.

4.1. Overview of road traffic safety methods and road safety analysis applications

4.1.1 Application of road safety methods

Figure 7 shows the five instruments of the road infrastructure safety management and their scope of application according to the number of accidents which have been occuring on the road network.

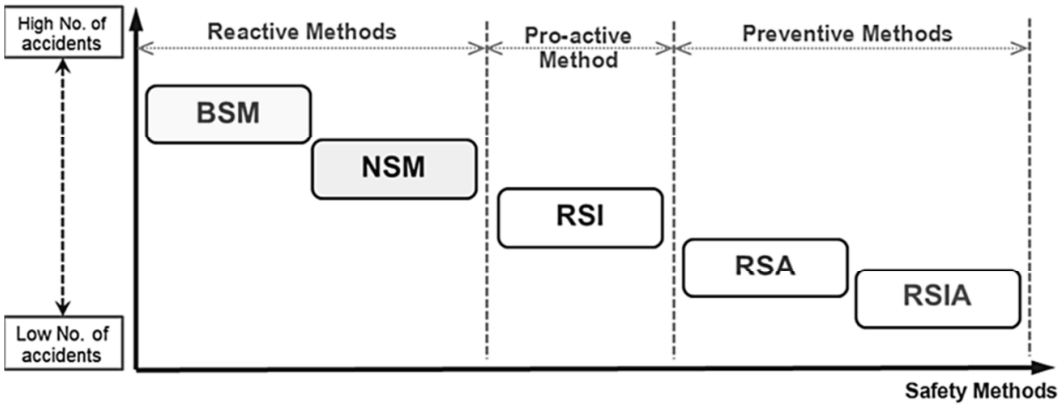


Figure 7. Application of road safety methods

CEDR (2008) claimed that the infrastructure safety management focuses on the five instruments: Road Safety Impact Assessment (RSIA), Road Safety Audits (RSA), Road Safety Inspections (RSI), Network Safety Management (NSM), and Black Spot Safety Management (BSM). These five instruments can be basically defined as follows.

The RSIA is a strategic comparative analysis of the impact of a new road or a substantial modification to the existing network on the safety performance of the road network, at the initial planning stage before the infrastructure project is approved. The RSA is an independent detailed systematic and technical safety check relating to the design characteristics of a road infrastructure project and covering all stages from planning to early operation as to identify, in a detailed way, unsafe features of a road infrastructure project. RSI is a systematic, periodic, objective and proactive safety review of a road in operation. The objectives of RSI are to identify and eliminate hazardous conditions, faults and deficiencies in order to improve the road safety for the road users. NSM is used to manage the existing road network or the parts of the road network with the aim to identify, localise and rank road sections according to their potential savings in accident costs. BSM is the reactive investigation and implementation of remedial measures at single localized (e.g. curves, junctions), short road segments or sites with a high number of injury accidents. BSM and NSM are reactive approaches to improve the safety performance of road infrastructure during operation.

4.1.2 Application of road safety analysis

Analysis of road safety is an important field of road engineering which helps develop strategies to reduce especially the number of fatalities in the future. Brannolte and

Munch (2009) pointed out four levels of safety analysis as shown in Figure 8. The process of these four safety analysis levels can be described as follows.

First, a general view on the whole road network will be provided by accident mapping analysis. Mapping accidents (i.e. location, categories, types, circumstances, road users, etc.) is an essential prerequisite for drawing sound conclusions with regard to accident countermeasures. This applies in particular to accident accumulation sites (black spots). The lowest level of safety analysis helps us to find out which particular locations in a big region should be considered to improve road traffic safety.

Second, macroscopic level of safety analysis (NSM) focuses on identifying, analyzing and classifying parts of the existing road network according to their potential for safety development and accident cost saving. It also helps the road administrations in detecting the sections within the network with the highest SAPO, i.e. where an improvement of the infrastructure is expected to be highly cost efficient. Suitable measures can then be derived from a comprehensive analysis of the accidents. The safety potential and the calculated cost of the measure together form the basis for an economic assessment, which is usually conducted as a cost – benefit analysis.

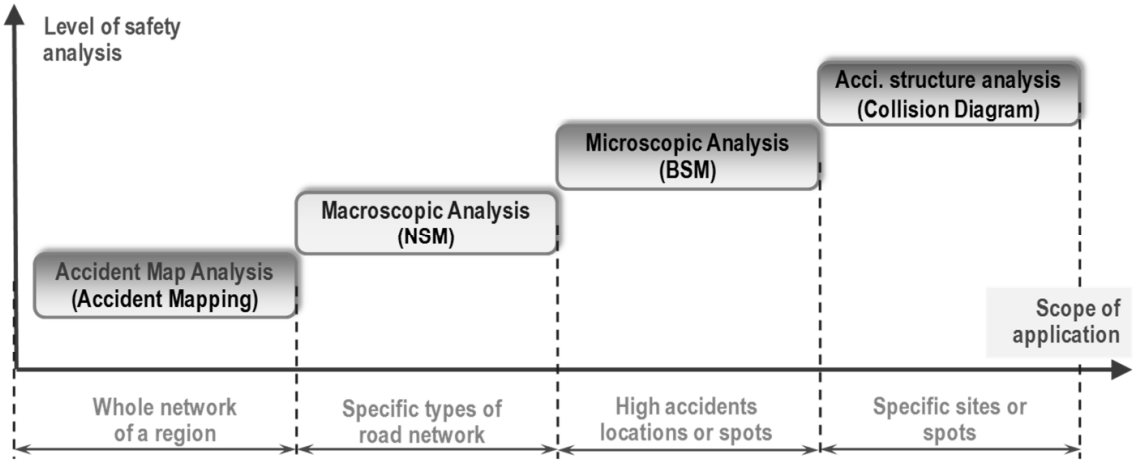


Figure 8. Road safety analysis applications

Third, BSM goes up to microscopic level of safety analysis with the aim to identify, analyze, and rank the high accident concentration sections or spots within the road network which have been in operation for more than three years and upon which a large number of fatal accidents in proportion to the traffic flow have occurred.

The highest level of safety analysis is the collision diagram analysis with focus on the concentrations and similarities of accidents at black spots selected. After black spots are identified, the accident data at those spots can be analyzed in order to find common patterns in accidents. A visit of the black spot site is usually part of the process of analysis. Collision diagram can be used in investigation of conflict situations on local spots. On the other hand, collision diagrams have been proven as very useful tools for detecting safety deficits easily.

4.2. Limitations of traditional BSM approaches

4.2.1 Limitations in identifying black spots

In order to identify black spots, traditional approaches only rely on the recorded number of accidents. Such dependence may pose the problem of inaccuracy in value because any value of recorded number of accidents at a site during a certain period is always a sum of two values: the first value is the very systematic number, and the second value is a random number. Accordingly, if a value of recorded number was used as a critical value in identifying black spots, then the identified black spots will be a set of both true and false black spots.

Indeed, Elvik (2008) used expected number of accidents concept and simulated data to point out the limitations and pitfalls of identifying road accident black spots in terms of the recorded number of accidents only. Based on the analysis we may now define four categories of sites as follows, assume that E denotes the expected number of accidents, [c] denotes the selected critical value, and R denotes the recorded number of accidents at a site during a given period of time.

① *Correct positives*: if $E \geq [c]$ and $R \geq [c]$

② *False positives*: if $E < [c]$ and $R \geq [c]$

③ *Correct negatives*: if $E < [c]$ and $R < [c]$

④ *False negatives*: if $E \geq [c]$ and $R < [c]$

Here, true and false are defined based on the comparison of expected number and selected critical value. Positive and negative are defined based on the comparison between values of recorded number and selected critical value. The expected number of accidents is the average number of accidents that will occur per unit of time in the long run, given that exposure and all risk factors remain constant.

"The performance of the various criterion values can be assessed quantitatively in terms of screening performance criteria developed in epidemiology (Deeks 2001, Rothman and Greenland 1998). Two of the most common criteria for diagnostic tests are sensitivity and specificity," (cited in Elvik, 2008, p. 26). They are defined as follows:

$$Sensitivity = \frac{Number\_of\_correct\_positives}{Total\_number\_of\_positives} \tag{11}$$

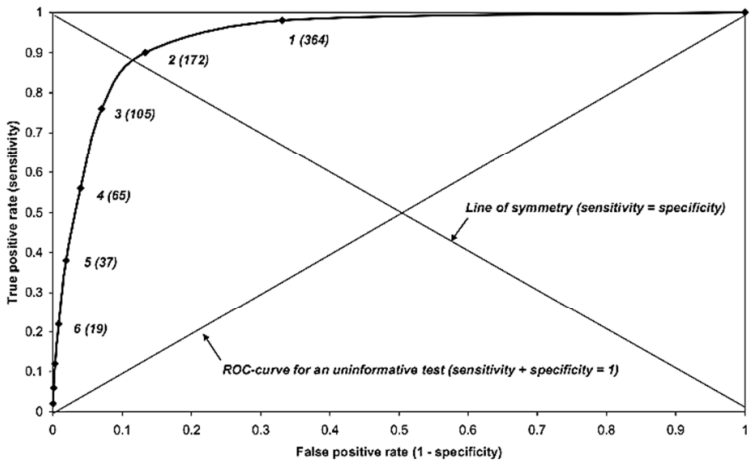
$$Specificity = \frac{Number\_of\_correct\_negatives}{Total\_number\_of\_negatives} \tag{12}$$

The total number of positives equals the number of correct positives plus the number of false negatives, and the total number of negatives equals the number of correct negatives plus the number of false positives.

The performance of different values for the critical number of accidents used to identify a black spot can now be assessed in terms of a ***receiver operating characteristic curve*** (ROC-curve). Such a curve, derived from the simulated data, is shown in Figure 9. The false positive rate is plotted along the abscissa. This is equal to 1 minus specificity. The true positive rate (sensitivity) is plotted on the ordinate. If the diagnostic

test discriminates well, the ROC-curve will rise steeply, close to the ordinate and flatten out near the top of the diagram. If the diagnostic test is uninformative, the ROC-curve will follow the diagonal line indicated in Figure 9.

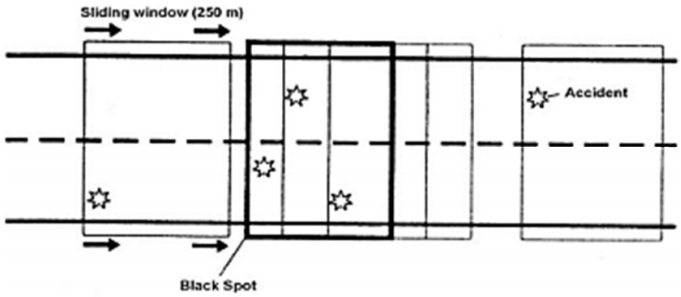
It is desirable to minimize the false positive rate and to maximize the true positive rate. This involves a trade-off; one may diminish the false positive rate by accepting a lower true positive rate, and vice versa. The optimal criterion is the one that maximizes the sum of sensitivity and specificity. For the Figure 9, this is to treat all sites with 2 or more accidents as potential black spots. This is marginally better than using 3 accidents as the criterion.



Source: TØI report 883, 2007

Figure 9. ROC-curve for detecting road accident black spots. Simulated data

It can be seen that no criterion for identifying hazardous road locations is perfect. The reason is very simple. We cannot observe the expected number of accidents. We can only observe the recorded number of accidents, which is always partly the outcome of chance, partly the outcome of very many factors that systematically influence the expected number of accidents. Thus, the choice of a criterion for identifying black spots cannot be based on a statistical criterion only. No statistical criterion can reliably identify only correct black spots, and include all of them, as the criterion would always be applied to a population of sites containing a mixture of random and systematic variation in the number of accidents. This mean that any criterion will be imperfect: Sites identified as black spots will always contain a mixture of correct positives and false positives. Besides, there will always be a number of false negatives. The number of sites that are true or false black spots will almost never be known.



Source: Austrian guidelines for black spot identification

Figure 10. Identification of road accident black spots in Austria by sliding window approach

With regard to the sliding window method, Elvik (2008) had two important conclusions. One is that using this method artificially inflates the number of black sections, and makes each section look blacker than it really is (i.e. having a higher recorded number of accidents); the other is that sliding window has the advantage of identifying more correct positives, but the disadvantage of identifying more false positives.

4.2.2 Limitations in analysis of black spots

Traditional approaches employ only recorded number of accidents to identify black spots. As a consequence, any set of identified hazardous road locations will contain both true and false positives. This may lead to the discrimination of true and false positive sites. This section summarizes some results from past studies in order to highlight the limitations in black spot analysis of traditional approaches.

Danielsson (1988) showed that "one commonly used criterion for identifying a truly hazardous road location, namely the over-representation of a particular type of accident is vulnerable to regression-to-the-mean-bias, because overrepresentation could be attributable mainly to chance," (cited in Elvik, 2008, p. 30).

"A commonly applied criterion to discriminate between true and false black spots is the presence of a dominant accident pattern. A dominant accident pattern is characterized by the overrepresentation of a particular type of accident," (cited in Elvik, 2008, p. 31). It is therefore of some interest to probe whether there is any difference in the regression-to-the-mean effect between hazardous road locations that have a dominant accident pattern and those that do not.

In order to test this, Elvik (2006) conducted a study on regression-to-the-mean at hazardous road sections with and without a dominant accident pattern and came to the conclusion that the presence of a clearly identifiable pattern of accidents characterized by the dominance of a particular type of accident may not effectively separate true from false black spots.

Harwood *et al.* (2002) pointed out that "some sites with a high number of accidents do not have readily identifiable accident pattern. A given deficiency in highway design or traffic control can contribute to accidents at one site, while at another site with similar deficiency, there are no accidents or no clear pattern of accidents associated with the deficiency," (cited in Elvik, 2008, p. 32). Finally, a given deficiency can contribute to different accident types. This suggests that an analysis of accidents designed to identify true black spots must go beyond merely identifying a dominant accident pattern.

Elvik (2008) concludes that "an approach to accident analysis is needed that provides clearer criteria for identifying true black spots, recognizes the possibility that analysis might be inconclusive, and minimizes the role of analyst expectancies" (p. 33).

4.3. Proposal of a new BSM approach for HCMC

4.3.1 Definition and Background

In the literature there is no universally accepted definition of a black spot. According to Elvik (2008) in practice there are three common types of black spot definitions as summarized in Table 1.

The numerical definition does not make any reference to traffic volume or to the normal number of accidents, nor does it specify the type of location considered. A statistical definition of an accident black spot relies on the comparison of the recorded number of accidents to a normal number for a similar type of location. Model-based definitions of road accident black spots are derived from a multivariate accident prediction model. Models were developed for intersections and road sections.

Table 1. Three common types of black spot definitions and descriptions

Numerical definitions	Statistical definitions	Model-based definitions
▪ Accident number	▪ Critical value of accident number	▪ Empirical Bayes
▪ Accident density	▪ Critical value of accident density	▪ Dispersion value
▪ Accident density & number	-	-

Virtisen (2002) stated that "high-risk sites are targeted with aim of improving safety on the road network through remedial treatment of the sites. Any achieved positive effects of safety measures at accident hot spots are denoted the benefits of the implemented measures. Implementing safety measures is costly, but in theory, all measures generating a positive net-benefit should be implemented. However, the restricted funding for hot spot safety work does put a limit to the number of sites that may be treated. Therefore, it is necessary to prioritize between sites and safety measures in order to utilize the limited funds as effectively as possible," (cited in Geurts and Wets, 2003, p. 11). The general aim of prioritizing may be described as:

$$\underset{Y}{Max}\ \frac{B(Y)}{C(Y)} \tag{3}$$

where,

Y represents a portfolio of safety measures, and C(Y) and B(Y) denote the corresponding overall cost and benefit of Y.

4.3.2 The aim of new BSM approach

The aim of new approach is to enable road administration to:

(1) Determine spots within the road network with a poor safety performance based on accident data and where deficits in road infrastructure have to be suspected;

(2) Rank the spots by potential savings (safety potential) in accident costs in order to provide a priority list of spots to be treated by road administrations.

The accident structures of the spots are then analyzed in order to detect abnormal accident patterns, which can lead to possible improvement measures. Finally, this offers

the possibility of comparing the costs of improvement measures to the potential savings in accident costs, allowing the ranking of measures by their cost–benefit ratio.

4.3.3 Advantage of the approach

The new BSM approach called SAPO-Based BSM developed is based on the *Network Safety Management* (BASt & Sétra, 2005). The methodology employed in the approach focuses on the traffic volume and the severity of accidents at spots and the evaluation of the accidents on the basis of accident cost rates. The comparison of actual accident cost with a hypothetical estimated base accident cost provides information on SAPO of spots.

The advantage of the SAPO compared to the classic accident parameters is that it allows assessing different road types and roads with different volumes at the same time. Furthermore, as the SAPO is given in accident cost, it can be related to the cost of the improvement measures.

Brannolte & Munch (2009) claimed that SAPO is the most important parameter to identify black spots on which safety improvement measures are expected to have the greatest effects.

4.3.4 Basic values for determination of SAPO

The development of the SAPO-Based BSM Approach is based on the calculation of accident rates and accident cost rates. Thereby, the following different calculation models should be distinguished:

- Sections with similar alignment: models for certain road sections;
- Transitions (single elements): models for spots.

4.3.4.1 Accident Cost

When analyzing accidents of different categories together, the numbers of accidents are weighted by the accident severity. Accident Costs (AC) are, therefore, used to describe the combined effect of number and severity of the accidents.

Annual Average Accident Cost (ACa) [USD/year] is calculated with the formula as follows.

$$AC_a(F+SI+LI+PDO) = \frac{A(F)\times MCA(F) + A(SI)\times MCA(SI) + A(LI)\times MCA(LI) + A(PDO)\times MCA(PDO)}{t} \quad (4)$$

where,

A is number of accidents; MCA is the mean cost per accident [USD/acci]; and t is the period of time under review [year].

Table 2 shows the mean cost per accident for four different levels of severity in the road network in Vietnam. These accident unit costs were calculated based on the Human Capital Method.

Table 2. Mean cost per accident for various severities (Price level 2008)

Severity description	Cost per accident [USD/acci]
Fatal (F)	31,777
Serious Injury (SI)	9,488
Light Injury (LI)	1,071
Property Damage Only (PDO)	354

Source: JICA & NTSC (2009), *Study on Master Plan of Road Traffic Safety in Vietnam up to 2020*

4.3.4.2 Accident Density

The density of accidents represents the incidence of accidents on a road section within a defined time period. The accident density makes it possible to determine areas with a significant number of accidents. They are calculated based on the number of accidents/casualties or accident costs.

- Accident Density (AD): average number of accidents or casualties within either a road section of one kilometre length or a defined spot within a defined time period (number/year).
- Accident Cost Density (ACD): average of the economic costs for either a road section of one kilometre length or a defined spot within a defined time period (costs/year).

Due to the mentioned definitions the density is calculated with the formulas in Table 3.

Table 3. Formulas of Accident Density (AD) and Accident Cost Density (ACD)

Spots	Sections
$ACD = AC/(1000 \cdot t)$ (5)	$ACD = AC/(1000 \cdot L \cdot t)$ (6)
$AD = nA/t$ (7)	$AD = nA/L \cdot t$ (8)

ACD = density of accident costs; AD = density of accident; nA = number of accidents; L = length of road section; t = time period.

4.3.4.3 Accident Rate

Accident Rates represent a road user's risk of being involved in an accident. Like the densities, rates are calculated for numbers of accidents as well as for accident costs.

- Accident Rate (AR): average number of accidents at a traffic volume of one million vehicles and one kilometre section length (for spots only one million vehicles)
- Accident Cost Rates (ACR): average of economic costs at a traffic volume of one million vehicles and one kilometre section length (for spots only one million vehicles).

The rates for number of accidents or rather for accident costs are calculated with formulas in the Table 4 as follows:

Table 4. Formulas of Accident Rate (AR) and Accident Cost Rate (ACR)

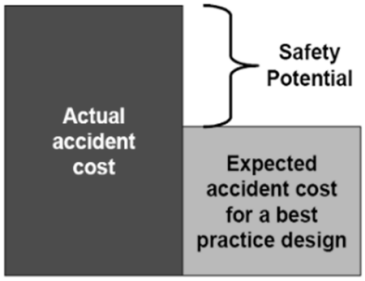
Spots	Sections
$ACR = 1000 \cdot AC/(365 \cdot AADT \cdot t)$ (9)	$ACR = 1000 \cdot AC/(365 \cdot AADT \cdot t \cdot L)$ (10)
$AR = 10^6 \cdot nA/(365 \cdot AADT \cdot t)$ (11)	$AR = 10^6 \cdot nA/(365 \cdot AADT \cdot t \cdot L)$ (12)

ACR = accident cost rate, AR = accident rate, AADT = average annual daily traffic, nA = number of accidents, L = length of road section, t = time period.

4.3.4.4 Safety Potential

It is an important task of road administrations to determine the road sections and spots that have poor safety properties which could be improved by changes in the roadway, its equipment, and traffic operation.

As resources are limited, those sections and spots where improvements can be expected to have the highest cost–benefit ratio have to be treated first. Therefore, information is needed on the accident costs per kilometer (or at a given location or spot) and the SAPOs for possible remedial measures.



Source: Transport Research Arena Europe 2006, Kerstin Lemke BASt, Germany

Figure 11. Diagrammatic views on the safety potential

The SAPO is defined as the amount of accident costs per kilometre road length (cost density) that could be reduced if a road section would have a best-practice design.

The higher the SAPO, the more societal benefits can be expected from improvements to the road. The SAPO (refers to Figure 11) is calculated as follows:

$$SAPO = ACD - bACD \tag{13}$$

The Basic Accident Cost Density (bACD) represents the anticipated average annual number and severity of road accidents per kilometre, which can be achieved by a best-practice design at the given average daily traffic. It can be calculated as the product of bACR and the Average Daily Traffic (ADT):

$$bACD = \frac{bACR \times ADT \times 365}{10^6} \tag{14}$$

The main idea is to define bACR for many different types of roads and intersections which are derived from the detailed assessment of existing accident cost rates. The

bACR includes only that share of all accidents which could not be avoided by a very good design on regulations conforming to standards.

4.3.5 Statistical Test

Table 5. Formulas for calculation of expected number of accidents (eA)

Spots	Sections
$eA = \frac{365 \cdot \overline{AR} \cdot ADT \cdot t}{10^6}$ (15)	$eA = \frac{365 \cdot \overline{AR} \cdot ADT \cdot L \cdot t}{10^6}$ (16)

$\overline{AR}$ = Average accident rate [A/(10^6 veh . km)], ADT = Average daily traffic in t years [veh/24h], L = Length of road section [km], t = Period of time under review [years].

According to *Network Safety Management* (BASt & Sétra, 2005) to make sure that the road sections or spots identified as hazardous are not merely the result of random variation in accident counts, statistical tests are performed. The test consists of the comparison of the observed number of accident A with the expected number of accidents eA of that section or spot and the determination of the importance of the deviation by calculating the confidence interval of the observed values (Poisson law).

4.3.6 Procedure of SAPO-Based BSM Approach

The SAPO-Based BSM Approach can be subdivided into five work steps as shown in the Table 6.

Table 6. Typical stages in SAPO-Based BSM Approach

Work Steps	Epigraphs	Descriptions
Basic data	Accident Pin Boards establishment	1-year & 3-year Accident Pin Boards (APBs) have to be established
1st	Identification of sites with high accident frequency	Identification of high accident sites based on the selected critical values and APBs
2nd	Ranking of sites	Ranking of high accident sites by SAPO
3rd	Accident pattern analysis and Local accident investigation	Collision diagram analysis and on-site inspection to determine safety countermeasures
4th	Treatment of identified BSs	Implementation of treatment including immediate, medium term & long term measures
5th	Evaluation	Before-and-after evaluation of effect of treatment

4.3.7 Required data

The following basic data are required for identification and treatment of black spots:

- Accident data (1-year and 3-year APBs refers to Figure 12)

- Geometric data of the road network
- Traffic volume of the road network

In the accident pin boards (APBs) accident type is marked by the colour of pins, accident category is marked by the size of pins, and some special accident circumstances are marked by coloured triangle such as pedestrians are involved, motorcycles are involved, alcohol, etc. as shown in the Figure 12.

- For each accident the following information is included:
  - Date and location;
  - Accident type, accident category, accident kind, and cause of accident;
  - Description of accident;
  - Number of injured persons (fatality, seriously injured, slightly injured);
  - Number of involved vehicles/pedestrians;
  - Further conditions: weather, light;
  - Influence of alcohol, drugs, medicine (driving while intoxicated).

Figure 12. Typical example of Accident Pin Board (APB) of the area under study

5. Conclusions

The traditional criterion for identifying a true black spot based on a dominant pattern of accidents has turned out to be of inadequate validity. Analysis of accidents at black spots is best viewed as a means of developing hypotheses regarding potential contributing factors to the accidents.

The new BSM approach is based on the safety potential in identifying and ranking black spots. The SAPO identifies network spots on which safety improvement measures are expected to have the greatest effect, but it requires a reliable basic data system. The basic data such as accident pin boards or accident maps, road network data, and traffic volume play the key role in identifying, ranking and treating black spots.

The new BSM approach is expected to fill the gaps left in traditional approaches in identifying and analyzing black spots; and to give satisfactory remedies for the problems of road safety suitable for the local traffic conditions of HCMC.

References

[1] AASHTO: *Highway Safety Manual – 1st Edition*, (2009).

[2] BASt and Sétra: *Network Safety Management – NSM*, Final version, (2005).

[3] Brannolte, U., Munch, A.: *Software-Based Road Safety Analysis in Germany*, in the 4th IRTAD CONFERENCE Proceeding, Seoul, Korea, (2009), p. 207–218.

[4] Conference of European Directors of Roads – CEDR: *Tools for Infrastructure Safety Management – Fact Sheets and Common Conclusions*, (2008).

[5] DIRECTIVE 2008/96/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL, 19 November 2008 on Road Infrastructure Safety Management, (2008).

[6] Elvik, R., Høye, A., Vaa, T., Sørensen, M.: *The Handbook of Road Safety Measures*, Emerald Group Publishing Limited, Howard House, Wagon Lane, Bingley BD16 1WA, UK, (2009).

[7] Elvik, R.: *State-of-The-Art Approaches to Road Accident Black Spot Management and Safety Analysis of Road Networks*, RIPCORD-ISEREST – WP6, (2008).

[8] Gatti, G., Polidori, C., Galvez, I., Mallschützke, K., Jorna, R., Van De Leur, M., Dietze, M., Ebersbach, D., Lippold, C., Weller, D., Wyczynski, A., Iman, F., Aydin, C.: *Safety Handbook for Secondary Roads*, RIPCORD-ISREST – WP13, (2007).

[9] Geurts, K., Wets, G.: *Black Spot Analysis Methods: Literature Review*, SteunpuntVerkeersveiligheid, (2003).

[10] Ganneau, F., and Lemke, K.: *Network Safety Management – From Case Study to Application*, Sétra, Bagneux Cedex, France & BASt, Bergisch Gladbach, Germany, (2006).

[11] Japan International Cooperation Agency - JICA & Vietnam National Traffic Safety Committee – NTSC: *Study on Master Plan of Road Traffic Safety in Vietnam up to 2020, Final report,* (2009).

[12] Nguyen, H. H., and Taneerananon, P.: *Reality of Urban Road Safety in Ho Chi Minh City and Suggested Solutions*, Journal of Society for Transportation and Traffic Studies (JSTS), Vol. 3, No. 1, (2012), pp. 9-21.

[13] Nguyen, H. H., and Taneerananon, P.: *Phenomenon of Black Spot Relocation in Ho Chi Minh City: Causes and Lessons*, Paper presented at the 8th Asia Pacific Conference on Transportation and the Environment – "Asian Transport Challenge" – APTE8 - 2012 Hat Yai, Thailand, June 6-8, (2012).

[14] Nguyen, H. H., and Taneerananon, P., Koren, C., Iamtrakul, P., and Vesper, A.: *Traditional Black Spot Safety Management Approaches: Potential Problems and Suggested Solutions*, Paper presented at the 18th National Convention on Civil Engineering – NCCE, Chiang Mai, Thailand, May 8-10, (2013).

[15] Robert, A., Tegge, Jang-Hyeon, J., Yanfeng, O.: *Development and application of Safety Performance Functions for Illinois*, Illinois Center for Transportation, (2010).

[16] Sørensen, M., Elvik, R.: *Black Spot Management and Safety Analysis of Road Networks - Best Practice Guidelines and Implementation Steps*, RIPCORD-ISEREST – WP6, (2008).

Inspection of barrelling of upset forgings based on digital photographs

F. Tancsics[2], E. Halbritter[1]

[1] University of Széchenyi István, Department of Material Science and Applied Technology,
9026 Győr, Egyetem tér 1.
halbritt@sze.hu
[2] RÁBA Axle Ltd., Production Development,
9027 Győr, Martin út 1.
ferenc.tancsics@raba.hu

Abstract: Majority of multi-cavity forging operations starts with descaling upset (e.g.: spindle) or with upsetting that results in significant deformation (e.g.: ring gear). During upsetting between parallel pressure plates the work-piece is getting barrelled. Barrelling is in close connection with friction coefficient. Till now the inspection of the shape of the work-piece getting barrelled was only limited to some references or to measurements of some work-pieces that were upset to barrel-shape [1] [2] [3]. More thorough observation requires application of statistical methods. We can obtain valuable information through statistics based on photographs taken during the production as well. Our present article reports on the methods we tried and on the results we obtained.

Keywords: material flow, upsetting, barrelling, friction coefficient

1. Introduction and targets

RÁBA Axle Ltd. produces majority of the forgings in multi-cavity die forging (Figure 1).

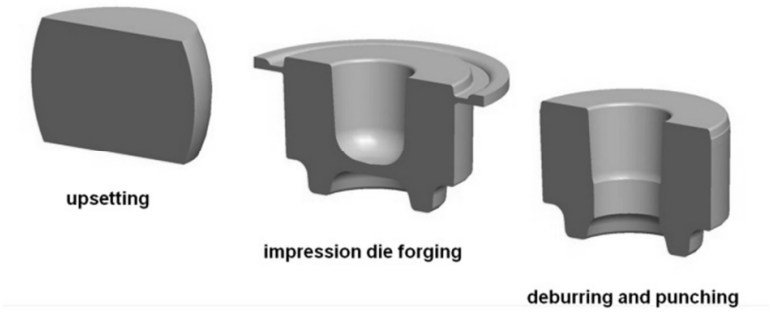


Figure 1: Phases of forming a planet gear forged in die

During upsetting between parallel pressure plates there is significant friction occurring at the connection of the work-piece to the pressure plates, which prevents free widening of the work-piece and consequently the surface of the work-piece is getting barrelled [2]. Barrelling shapes can only be observed in an intermediate state of the

work-piece. Intermediate shapes can be observed and evaluated by interrupting the production process or without interruption. As the intermediate shape can be characterized with statistical methods only, production of huge number of intermediate parts would be a perfect solution for getting information; but it is very expensive. However, based on photographs taken during production we can gain information about the actual shape of the upset work-piece. In our articles published earlier [1] [3] we elaborated a simplified method for determining friction coefficient based on deformation of a solid cylindrical body that was upset between parallel pressure plates by using a kinematically admissible velocity field. Our method assumes that the profile curve of the barrelling surface can be well approximated with a second-order polynomial. In order to check the profile curves of the work-pieces [1] [3], earlier we scanned the surfaces of the upset work-pieces with a GOM (*General Outcome Measurement)* optical digitizer owned by the University of Széchenyi István (Figure 2).

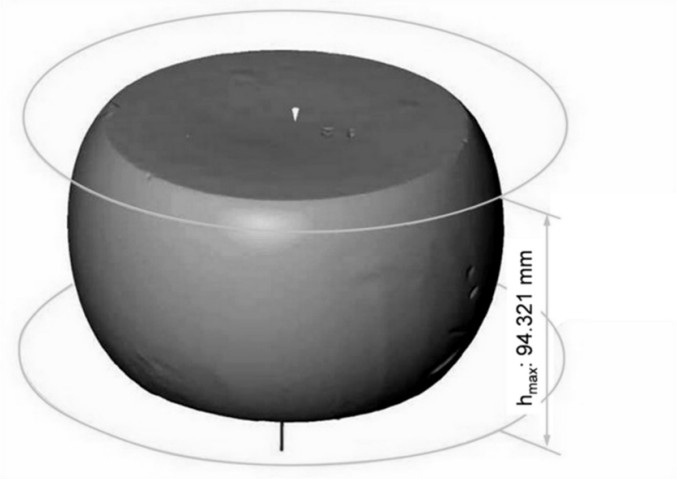


Figure 2: Optically digitized picture of an upset work-piece

The digitized surface was cut, coordinates of the points belonging to the selected part of the sectional boundary curve were saved, and then the points were approximated by second-order polynomial, using MathCAD software [1] [3].

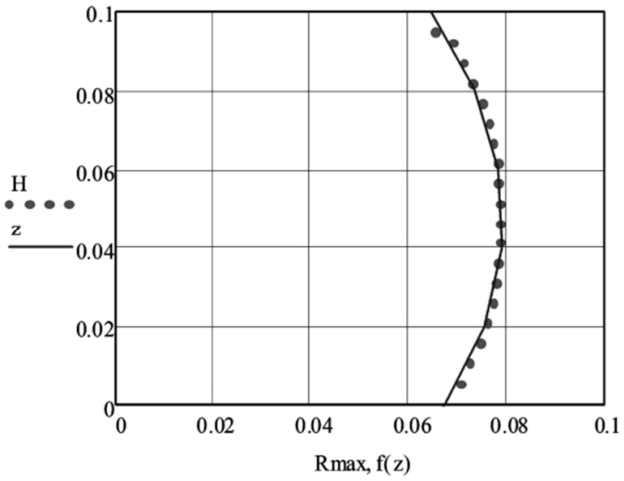


Figure 3: Curve of approximate function and measured points /upset height h=95 mm/

Correlation coefficient $r=0.996$ showed close connection between the values produced by the approximate function and the measured data (Figure 3).

Optical digitalization has the advantage that the work-piece can be scanned with no distortion, and based on the three-dimension point distribution the profile curve can be inspected for as much sections as desired. Disadvantage of this method is that the part inspected cannot be put back into the production process, the upset part can only be used as a scrap.

It has became necessary to select and develop a method, a cost-saving procedure, which can provide observation of the work-piece upset to barrel-shape during continuous production and which can statistically prove how much approximation of the profile curve can be accepted by a second-order polynomial in case of a barrelling work-piece and how much application of the second-order polynomial can be generalized.

2. Evaluation on the basis of digital photos

As it was targeted, shape of the upset work-piece had to be observed under production circumstances. In case of the multi-cavity forging technologies applied by the Forging Plant of Rába Axle Ltd., after pre-forming the upset work-piece is immediately put into a die cavity on the same forging machine, so this instantaneous period between two forming can be used for observations. So the applicable methods are largely limited. If we want to be real, only one photograph can be taken during the available period of time. It helps to take a photo that the hot steel has typical colour at the forging temperature. This typical colour can be more or less distinguished from the colour of the environment, no special light is needed [8] (Figure 4).

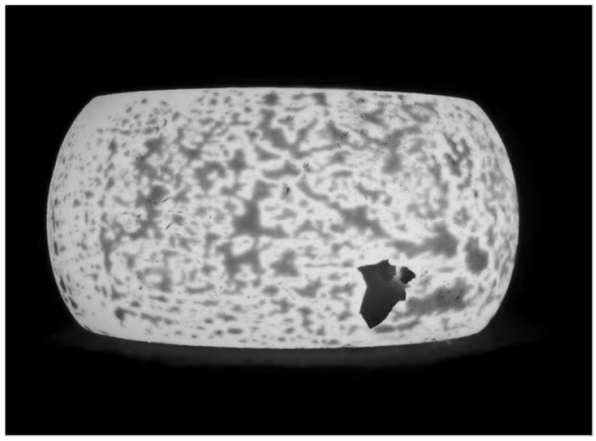


Figure 4: Photo of the upset work-piece at forging temperature

This photo means scanning with narrowed information. Projected picture of the work-piece was used in the evaluation. Projection lines are not parallel, so in the picture of the upset work-piece the boundary line only approximates the profile curve [5]. In our opinion, this is an acceptable approximation for the inspection.

The photo scans the shape of the work-piece from a single aspect, but if we assume that the barrelling work-piece is axially symmetrical, then this narrowing is also acceptable [9]. The digital photo was evaluated with different methods.

2.1. Evaluation based on pixels

Evaluation of photo-like pictures is widely used in various areas of industry [5] [6] [7]. The method of our evaluation is original as the task has been solved with available means, under conditions of hot forming, without interrupting the production process.

MathCAD mathematical software supports processing of the pixels to some extent. The colour picture /Figure 4/ taken on the work-piece at forging temperature can be read, reduced or enlarged.

In the evaluation, not the dominating wavelength of pixels (tint) or common value of colour saturation (chromaticity) are important, but that the work-piece shall be well distinguished from its environment. Gray-scale picture simplifies picture processing, as only one colour code belongs to a pixel. Colour codes are varying between 0 and 255. 0 corresponds to black and 255 to white. MathCAD software stores the codes relating to the pixels in a matrix. Code of the pixels can be replaced in the matrix.

First, we made the picture binary, i.e.: the code of every pixel was changed to 1 within the boundary line of the work-piece, and to 0 outside of the boundary line. To make this change, we had to assign a number /code/ between 0 and 255, which could separate the work-piece from the environment [6].

To look at the condition after filtering, only the matrix containing 0 and 1 was multiplied with 255. After this operation was made, we got the picture shown in Figure 5.

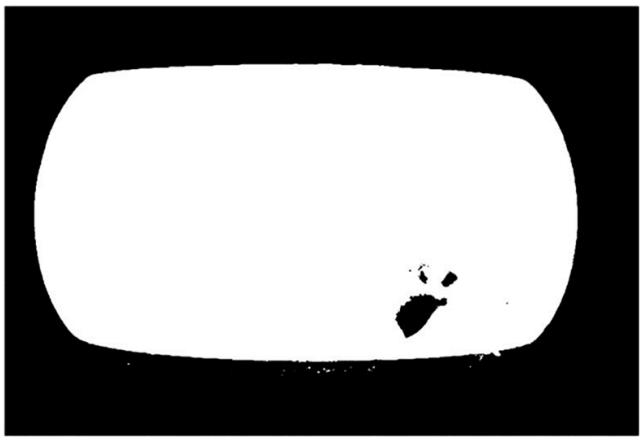


Figure 5: Result of primary filtering

Based on the Figure 5 it can be established that this primary filtering is not perfect, the dark scale spots in the bright field as well as the white points at the strong reflection locations disturb the evaluation of the picture, contour of the work-piece is “noisy” at some locations.

At the wrong locations, secondary filtering was made to improve the matrix values. By means of the *Picture Toolbar* we encircled the spot we wanted to remove and created a partial matrix. By setting the proper colour codes of the partial matrix, we can modify the selected spot in the original matrix. After the black spots inside the work-piece were removed, we got the picture shown in the Figure 6.

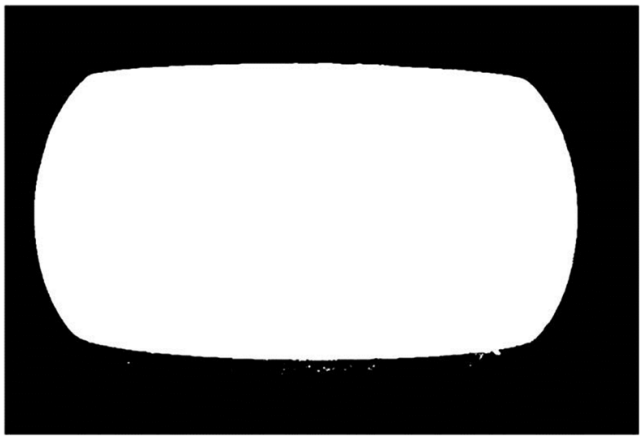


Figure 6: Results of secondary filtering

Depending on the target of the evaluation, filtering the white spots outside of the work-piece could be necessary. Filtering resulted in locating 1 at the location of the white colour and 0 at the location of the black colour in the matrix corresponding to the pixels. Numbers and distribution of the pixels on the work-piece can be detected by elaborating the matrix. After the matrix is evaluated, the barrelling shape can only be retrieved, if the evaluation is made to the half of the picture (Figure 7).

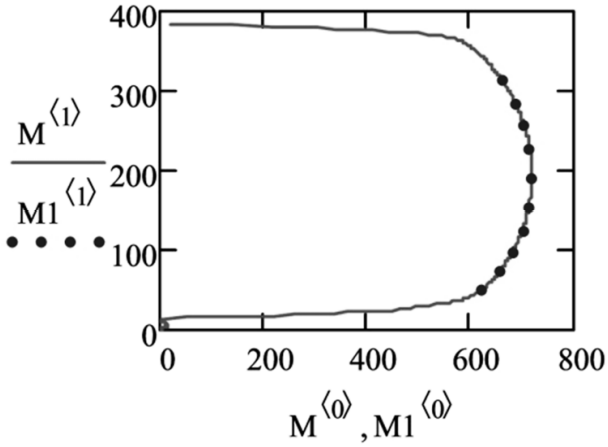


Figure 7: Plotting the boundary curve based on pixels (contour line of the work-piece is indicated with continuous line, the approximate curve is indicated with dots)

Points on the profile curve of the work-piece can be separated from the matrix, based on the separated points the coefficients of the second-order polynomial can be determined with regression calculation, correlation calculation can detect whether the second-order approximation is proper.

After regression calculation and correlation calculation are made with MathCAD software for the case shown in the Figure 7, the correlation coefficient is $c=0.998$, which refers to a close connection.

On the lower part of the Figure 7 it can be observed that under the boundary curve there is a small red spot near the zero. This is caused by the fact that the white spots outside of the work-piece were not filtered. In the evaluation presented in the Figure 7 there are maximal 384 pixels vertically and 719 horizontally.

2.2. Evaluation with CAD software

AutoCAD software supports loading of raster images */Insert ► Raster Image Reference…/*. The loaded image corresponds to a background image, but based on it the boundary points can be selected (Figure 8).

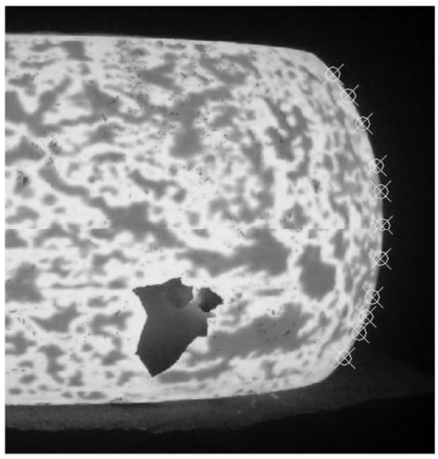


Figure 8: Selection of boundary points in AutoCAD environment

It makes easier to select the boundary points that the loaded image can be enlarged as desired. For later data processing it is practical to collect and save the coordinates of the selected points. Software was written to record and save the data in programming language AutoLISP. The saved values related to the Figure 8 are shown in the table No.1.

Table 1: Coordinates of the selected boundary points

Serial number	R (radial direction)	z (axial direction)
0	34.325	23.073
1	34.764	22.453
2	35.181	21.643
3	35.575	20.428
4	35.706	19.689
5	35.758	18.692
6	35.701	17.678
7	35.463	16.563
8	35.273	15.995
9	35.076	15.536
10	34.651	14.809

In the Table 1 the values are given in a cylinder coordinate system where the column *R* indicates the dimensions in radial direction and the column *z* shows the dimensions in axial direction. Saved values were processed with MathCAD software. The function approximating the points was also defined with MathCAD software and it was used to

inspect how close connection the values given by approximate function and the measured values have. After correction calculation was made with the data shown in the table 1, the correlation coefficient is *c=0.999*. Close connections can be well detected based on the points and the plot of the function approximating the points, too (Figure 9).

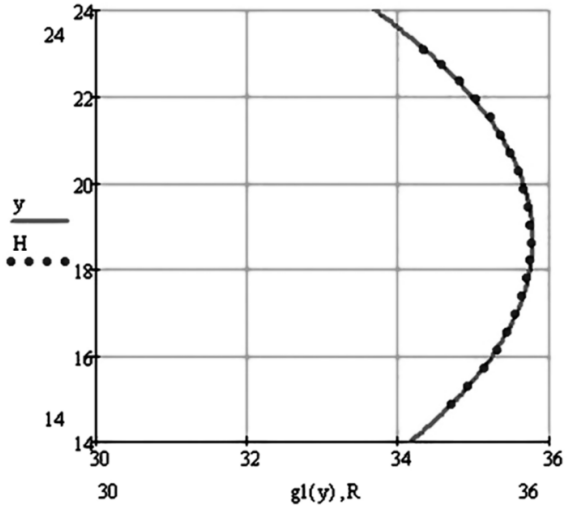


Figure 9: Selection of boundary points in AutoCAD environment

It can be established that two totally different methods for evaluation of images provided practically the same results.

So the barrelling profile curve can be well approximated with a second-order polynomial.

We have established that none of the evaluation methods can completely be automated, both require individual decisions. In the evaluation based on pixel numbers, some additional issues can be caused by scale appearing at unpredictable locations. Calculating with the probable difficulties in the evaluation, we think it is more favourable to apply AutoCAD software together with AutoLISP.

3. Inspection of samples taken out of production

In order to prove our theory, we have made some experiments in Rába Forging Plant under constant production conditions. During the experiments, at various L/D (*Length/Diameter*) conditions digital photos have been taken on work-pieces that were upset to different heights and extents. Photos have been taken with fixed camera, with flash, at the same distance and position. Setting data of the camera are summarized in the table No.2. When the settings needed for taking the photos were selected, we intended to reduce the perspective distortion and the distortion effects of the lens [5] [6]. Using the recommended method of AutoCAD/AutoLISP, we have checked how the profile curves can be approximated with a second-order polynomial. Acceptable approximation was defined as the connections that are closer than the correlation coefficient *c=0.995* given in the reference [4].

Typical parameters and data:

- cutting length: H_0 (mm),
- initial diameter: D_0 (mm),

- upsetting temperature: $T_{upsetting}$ (°C),
- upset height in cooled condition: h_n (mm),
- upset largest diameter in cold condition: D_{max} (mm),
- lubrication: without lubrication,
- machine: 40 MN MAXIMA crank press,
- surface roughness of die surfaces: Ra=0.25,
- pixel width x pixel height: 4288 x 2848,
- resolution (*DPI = Dots Per Inch*): 300,
- focal distance: f (mm),
- exposure time: exp (sec),
- ISO sensitivity: 200,
- flash: used.

Table 2: Typical data of upset work-pieces

Upset work-piece No.1.							
Forging					Photo		
H_0	D_0	$T_{upsetting}$	h_n	D_{max}	f-stop	exp	f
150	**90**	**1130**	**128.1**	**99.8**	**f/8**	**1/200**	**52**
Upset work-piece No.5.							
Forging					Photo		
H_0	D_0	$T_{upsetting}$	h_n	D_{max}	f-stop	exp	f
150	**90**	**1150**	**114.8**	**107**	**f/10**	**1/125**	**52**
Upset work-piece No.7.							
Forging					Photo		
H_0	D_0	$T_{upsetting}$	h_n	D_{max}	f-stop	exp	f
100	**90**	**1110**	**65.8**	**118.2**	**f/10**	**1/60**	**50**
Upset work-piece No.10.							
Forging					Photo		
H_0	D_0	$T_{upsetting}$	h_n	D_{max}	f-stop	exp	f
100	**90**	**1158**	**28.8**	**175.1**	**f/10**	**1/60**	**50**
Upset work-piece No.14.							
Forging					Photo		
H_0	D_0	$T_{upsetting}$	h_n	D_{max}	f-stop	exp	f
150	**90**	**1147**	**78.2**	**131.9**	**f/10**	**1/80**	**50**

Geometry of the upset work-piece No.1. is typical to descaling upset.

During descaling upset, forming is only made to an extent so that the scale layer on the work-piece surface shall be disturbed enabling the removal of the scale (Figure 10).

Figure 10: Geometry of the sample No.1. at 14.6 % height reduction. Correlation coefficient: c=0.997

Based on the table, correlation coefficient of the sample No.5. shows good approximation, but does not meet the requirements established. Geometry of the work-piece is presented in the Figure 11.

Figure 11: Geometry of the sample No.5. at 23.45 % height reduction. Correlation coefficient: c=0.993

Upset profile shows visible distortion, which is resulted from inhomogeneous temperature distribution within the work-piece. In this case some intervention has become necessary. Following the intervention, again, proper profile curves were produced (Figure 12, Figure 13 and Figure 14). Value of the correlation coefficients was set to *c=0.998*, which shows very convincing approximation.

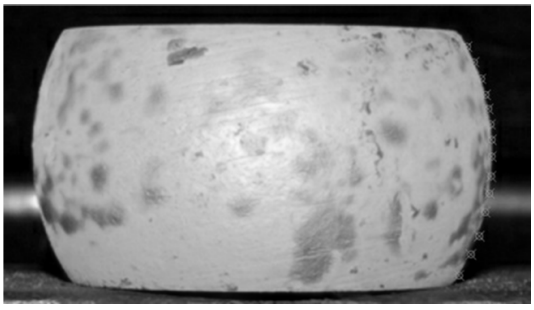


Figure 12: Geometry of the sample No.7. at 34.2 % height reduction. Correlation coefficient: c=0.998

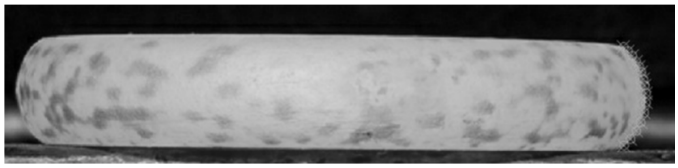


Figure 13: Geometry of the sample No.10. at 71.2 % height reduction. Correlation coefficient: c=0.998

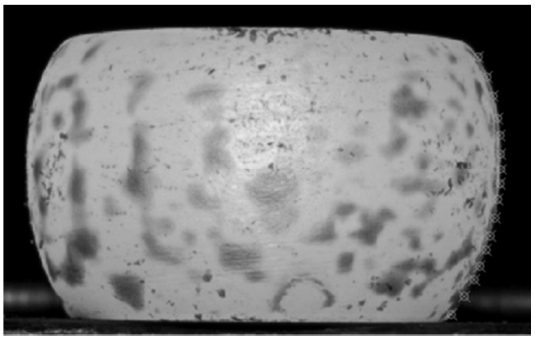


Figure 14: Geometry of the sample No.14. at 47.9 % height reduction. Correlation coefficient: c=0.998

Upset work-pieces No.7., No.10. and No.14. are geometries typical to pre-forming upset, where proper profile curves are highly expected.

4. Conclusions and Future Improvements

During our inspections our primary aim was to observe characteristics of barrelling, to investigate how much approximation of the profile curve can be accepted with a second-order polynomial, and how much the second-order polynomial can be generalized in a particular area.

Accuracy of approximation was inspected with significantly different methods. The results we obtained showed significant equivalence, confirmed our assumption that the upset profile curves can be well approximated with a second-order polynomial. In our inspections, pressure plates of constant surface roughness and lubrication-free technology were applied. Similar production circumstances also enabled us to identify the determinant parameters that can be related to distortions of the upset profile curve. We have established that goodness of the approximation barely depends on the extent of upsetting and L/D relation, but significantly depends on internal heat distribution of the work-piece.

Based on the results obtained, we can surely declare that in hot-forming, barrelling of the work-piece can be approximated with a second-order polynomial with proper accuracy if heat distribution of the work-piece is homogeneous.

The results are exceeding the inspections of the profile curve, as development of the recommended method AutoCAD/AutoLISP could facilitate direct control of the significant geometrical parameters (shape, dimension) of the pre-formed work-pieces in hot condition. All we have to do is to attach the camera to a computer system and to complete algorithmization of the presented control method according to the task.

References

[1] Kozma, I., Tancsics, F., Halbritter, E.: *Modelling of the Expectable Shape of the Barrelling Working Piece*, OGÉT 2010 18th International Conference on Mechanical Engineering, Nagybánya, (2010), pp. 261-264.

[2] Tancsics, F., Halbritter, E.: *A súrlódási tényező újszerű meghatározása és felhasználása a Pro/Engineer és MathCAD szoftverek segítségével*, GÉP LXI/7, (2010), pp. 34-42 (in Hungarian).

[3] Tancsics, F., Halbritter, E., Kiss B.: *Simplified Determination of Friction Coefficient by Upsetting*, OGÉT 2009 17th International Conference on Mechanical Engineering, Gyergyószentmiklós, (2009), pp. 384-387.

[4] Tancsics, F., Kiss, B., Halbritter, E.: *Limit Analysis of Adaptation of the Mathematical Model Made to Determine Friction Coefficient,* OGÉT 2011 19th International Conference on Mechanical Engineering, Csíksomlyó, (2011), pp. 355-359.

[5] Cintrón, R., Saouma, V.: *Strain Measurements with the Digital Image Correlation System Vic-2D*, Department of Civil Environmental and Architectural Engineering University of Colorado, (2008).

[6] Pickle, J.: *Measuring the Area of Irregular Shaped Objects in Digital Images Using Image Analysis Software,* Museum of Science, (2005).

[7] Kornis, J.: *Optikai mérések alkalmazása járműipari mérésekben*, Budapesti Műszaki és Gazdaságtudományi Egyetem Fizika Tanszék, (2010) (in Hungarian).

[8] Marschner, S.R., Westin, S.H., Lafortune, E.P.F., Torrance, K.E., Greenberg, D.P.: *Image-Based BRDF Measurement Including Human Skin*, Program of Computer Graphics Cornell University, (1999).

[9] Peterson, K.A.: *Introduction to Basic Measures of a Digital Image for Pictorial Collections*, Library of Congress, Washington, (2005) ps. 8.

Nonlinear Explicit Method for First Order Initial Value Problems

P. K. Pandey *

Department of Mathematics
Dyal Singh College (Univ. of Delhi)
Lodhi Road, New Delhi -110003, India
e-mail: pramod_10p@hotmail.com

Abstract: In this research paper, we present the development, implementation and analysis of a nonlinear explicit numerical method based on local assumption for solving first order initial value problems in ordinary differential equations. The method has at least second order accuracy and A-stable property. The property of A –stability suggests that method can be used for the solution of stiff initial value problems in ordinary differential equations. The method applied to find the numerical solution of several model problems. The computational results obtained for the model problems suggest that method is accurate, reliable and ingenious. The computational performance of the present method is comparable to the other method.

Keywords: Nonlinear method, Explicit method, Initial value problems, A -stable method.

AMS 2000 Subject Classification: 65 L 10, 65 L 12.

1. Introduction

Many problems that arise in natural sciences or social sciences, modeled mathematically in form of ordinary differential equations. For example decay of radioactive substances, economic growth, climatic change or logistic support distribution, these model problems in general have following form as initial value problems in ordinary differential equations

$$y'(x) = f(x,y) \quad , \qquad y(a) = y_0 , x \in [a,b], y, f(x,y) \in \mathbb{R} \tag{1}$$

However it is disappointing if we think about solution of these problems. Relatively few differential equations have analytical solutions .Since analytical solutions in general do not exist for these problems, an approximate solutions have to be obtained by numerical methods. We shall be concerned in this article, a numerical method for solving those initial value problems whose solutions can not be obtained analytically. In solving these problems, approximations plays an important role in development of algorithm.

Approximating the derivative by discrete expressions, the solution of this initial value problem can be obtained by moving away from the specified initial condition. We propose a numerical method that is efficient and reliable for solving stiff initial value problems of the form (1). We shall assume that (1) is well posed with continuous derivatives and that the solution depends differentially on the initial condition. This article is organized in six sections. Section 2 deals with development of the method while local truncation error estimated in section 3. The convergence and stability analysis of the method, discussed in sections 4 and 5. Numerical experiments on model test problems discussed in final section 6.

2. Description of the method

We define the mesh points of the interval $[a, b]$ in the usual way ,

$$x_i = a + ih \quad , \quad i = 0,1,2, \dots \dots \dots n \, , \tag{2}$$

being h- step size and $x_n = b$. Let y_i represent an approximate value of the theoretical solution $y(x)$ at the mesh point $x = x_i \; and \; f_i$ represent $f(x_i \, , y_i)$. Further we assume that $\frac{\partial f}{\partial y}$ is continuous in $[a, b]$, so (1.1) possess a unique solution [4].

Suppose we have solved numerically the problem (1) up to mesh point x_i and obtained a numerical value y_i as an approximation of $y(x)$ at $= x_i$. Let us assume local hypothesis [4] that $y(x_i) = y_i$, we are interested in obtaining an approximate value y_{i+1} for exact value of $y(x)$ at $x = x_{i+1}$ .Following the ideas in [3,4] ,we propose an approximation to the theoretical solution $y(x_i + h)$ of problem (1) given by

$$y(x_i + h) = y(x_i) + a_0 \, h \, y'(x_i) \, e^{\{a_1 \, h \, y''(x_i)/y'(x_i)\}} \tag{3}$$

where a_0 and a_1 are real undetermined coefficients. Further it is assumed that x_i is not extreme point of $y(x)$ for any $i = 0,1,2, \dots \dots n$,otherwise $y'(x_i) = 0$.

From (3) ,we have

$$F_i(h, x, y, y', y'') = y(x_i + h) - y(x_i) - a_0 \, h \, y'(x_i) \, e^{\{a_1 \, h \, y''(x_i)/y'(x_i)\}} = 0$$

If we expand $y(x_i + h)$ in Taylor's series about mesh point $x = x_i$, we have

$$F_i(h, x, y, y', y'') = h \, y'(x_i) \, (1 - a_0 \, e^{\{a_1 \, h \, y''(x_i)/y'(x_i)\}}) + \frac{h^2}{2} y''(x_i)$$

$$= (1 - a_0 \,) h \, y'(x_i) + \left(\frac{1}{2} - a_0 \, a_1\right) h^2 \, y''(x_i) + O(h^3) \tag{4}$$

To determine unknown coefficients a_0 *and* a_1 , assume that coefficients of h *and* h^2 in (4) vanish. Thus we obtain a system of nonlinear equations

$$1 - a_0 = 0$$

$$\frac{1}{2} - a_0 a_1 = 0 \tag{5}$$

Hence ,we have

$$(a_0 , a_1) = \left(1 , \frac{1}{2}\right) \tag{6}$$

Substituting the values of (a_0 , a_1) from (6) in (3) ,we will get

$$y(x_i + h) = y(x_i) + h\, y'(x_i)\, e^{\{h\, y''(x_i)/(2\, y'(x_i))\}} \tag{7}$$

Using (1) and notations defined above, we can write our explicit method as

$$y(x_i + h) = y(x_i) + h\, f_i\, e^{\,h\, f_i'/(2\, f_i)} \tag{8}$$

where $f_i' = \left(\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}\frac{dy}{dx}\right)_{(x_i , y_i)}$

Thus we have developed an exponential single step method of the form $y_{i+1} = y_i + h\, G(h, y', y'')$, where G is an increment function depends on h , y' *and* y''**.** Repeating the procedure along the mesh points of the interval for discrete solution of the problem (1).

3. Local Truncation Error

The local truncation error at the mesh point $x = x_{i+1}$ in the method (8) using the exact arithmetic, is given by

$$T_{i+1} = y(x_{i+1}) - y_{i+1}$$

$$= y(x_{i+1}) - y(x_i) - h\, y'(x_i)\, e^{\{h\, y''(x_i)/(2\, y'(x_i))\}}$$

$$= \frac{h^3}{24}\left(4y^{(3)}(\xi) - 3\frac{(y_i'')^2}{y_i'}\right) + O(h^4) \tag{9}$$

where $x_i < \xi < x_{i+1}$. If we denote $\underset{a \le x_i \le b}{max} |T_{i+1}| = T$, and $\underset{a \le x \le b}{max} \left|y^{(3)}(x)\right| = M$,so

$$T \leq \frac{h^3}{6} M \tag{10}$$

Thus the order of local truncation error is $O(h^3)$.

4. Convergence Analysis

To include the effect of the rounding errors following the idea in [7], we introduce e a new approximation $\bar{y}_i$, which is defined by the same procedure ,except that rounding errors are allowed. We have

$$\bar{y}_{i+1} = \bar{y}_i + h\, \bar{y}_i' \exp\left(\frac{h\, \bar{y}_i''}{2\, \bar{y}_i'}\right) - R_{i+1} \tag{11}$$

where $\bar{y}_i' = f(x_i, \bar{y}_i)$, $\bar{y}_i'' = \frac{\partial f(x_i, \bar{y}_i)}{\partial x}$ and the rounding error R_{i+1} is the amount by which the method (8) is not satisfied by $\bar{y}_i$. Applying (8) and (11) to test equation

$$y' = \lambda y \ ,$$

$where\ \lambda = \left(\frac{\partial f}{\partial y}\right)$ at some point $x_i \in [a, b]$ and subtracting ,we have

$$y(x_{i+1}) - \bar{y}_{i+1} = y(x_i) - \bar{y}_i + \lambda h(y_i - \bar{y}_i) \exp\left(\frac{\lambda h}{2}\right) + T_{i+1} + \mathrm{R}_{i+1} \tag{12}$$

If we define $e_i = y(x_i) - \bar{y}_i$,the error at $x = x_i$ and substitute in (12) ,we have

$$e_{i+1} = e_i + \lambda h\, e^{\left(\frac{h\lambda}{2}\right)} e_i + T_{i+1} + \mathrm{R}_{i+1}$$

$$e_{i+1} = \left(1 + \lambda h + \frac{(\lambda h)^2}{2}\right) e_i + T_{i+1} + \mathrm{R}_{i+1} \tag{13}$$

Let introduce the difference equation, so (13) can be written as

$$E_{i+1} = AE_i + B \tag{14}$$

where $\left|1 + \lambda h + \frac{(\lambda h)^2}{2}\right| \leq A$ and $|T_{i+1} + \mathrm{R}_{i+1}| \leq B$. It is clear $|e_i| \leq E_i$ if $|e_0| \leq E_0$, we obtain

$$E_i = A^i E_0 + \left(\frac{A^i - 1}{A - 1}\right) B, \ \ provided \ \ A \neq 1 \qquad (15)$$

Substituting $E_0 = |e_0|$ and $\left|1 + \lambda h + \frac{(\lambda h)^2}{2}\right| < e^{\lambda h}$ in (15), we have

$$|e_i| < \exp\big(\lambda(b-a)\big)|e_0| + \frac{\exp\big(\lambda(b-a)\big) - 1}{\lambda h + \frac{(\lambda h)^2}{2}}(T + R) \qquad (16)$$

where $R = \max_{a \le x_i \le b} |R_{i+1}|$.

If $e_0 = 0 \ and \ R_0 = 0$,i.e. rounding error is negligible

$$|e_i| < \frac{\exp\big(\lambda(b-a)\big) - 1}{\lambda h(1 + \frac{\lambda h}{2})} T \qquad (17)$$

Then from (10) ,we have

$$|e_i| < O(h^2) M \ \frac{\exp\big(\lambda(b-a)\big) - 1}{\lambda(1 + \frac{\lambda h}{2})} \qquad (18)$$

So $|e_i| \to 0 \ as \ h \to 0$. So method (8) is convergent.

If $e_0 = 0 \ and \ R_0 \neq 0$,then

$$|e_i| < \left(\frac{T}{h} + \frac{R}{h}\right) \frac{\exp\big(\lambda(b-a)\big) - 1}{\lambda(1 + \frac{\lambda h}{2})} \qquad (19)$$

Since $T = O(h^3)$, we see bound will decrease as *h* decrease , until the contribution due to *R* becomes dominant ,at which further decrease in *h* will increase the bound.

5. Stability Analysis

If we apply (8) to solve the test equation $y' = \lambda y$, we have

$$y_{i+1} = y_i + h\lambda y_i e^{\left(\frac{h\lambda}{2}\right)}$$

$$\leq y_i \left(1 + \lambda h + \frac{(\lambda h)^2}{2}\right)$$

$$< y_i e^{\lambda h}$$

$$= E(\lambda h) y_i \qquad (20)$$

where $E(\lambda h)$ is second order approximation to $e^{\lambda h}$ .Thus method (8) is absolutely stable if $|E(\lambda h)| \leq 1$. Thus interval of stability of method (8) is $(-\infty, 0] \cup [\frac{4}{h}, \infty)$.

6. Numerical Experiments

In this section, we have reported the computational performance of our method (2.7) on several stiff initial value problems in ordinary differential equations. We have computed maximum absolute error on the mesh points in the interval of integration for these problems and have shown in the table for different values of N the number of mesh points. Let y_i be the approximate value of the theoretical solution $y(x)$ of a problem calculated by method (8) at mesh point $x = x_i$ .We have calculated maximum absolute error in solution of problem by

$$MAE\ (y) = \max_{a \leq x_i \leq b} |y(x_i) - y_i| \quad , \quad i = 1,2, \ldots\ldots\ldots\ldots, N$$

We have also calculated error in solution of problem at end point of the interval by

$$ERR(y(b)) = |y(b = x_n) - y_n|$$

For comparison MAE , ERR is calculated by method in [5] and presented in same table for the respective model problem. All computations in the examples consider were performed in the GNU FORTRAN environment version -99 compiler(2.95 of gcc) running on a MS Window 2000 professional operating system.

Example 6.1. A stiff problem taken from [1]**,**

$$y'(x) = -100\ y(x) + 99\ e^{2x} \quad , \qquad y(0) = 0$$

which has exact solution $y(x) = \frac{33}{34}(e^{2x} - e^{-100x})$, *on the interval* [0,1]. The MAE, ERR presented in Table 1.

Example 6.2. When special initial condition are applied to the logistic model, we obtain the logistic problem [6]**,**

$$y'(x) = y(x)(1 - y(x)) \quad , \qquad y(0) = 0.5$$

which has exact solution $y(x) = (1 + e^{x})^{-1}$, *on the interval* [0,1]. The MAE, ERR presented in Table 2.

Example 6.3. Let be the stiff system taken from [8]**,**

$$y'(x) = -1002\ y(x) + 1000\ \left(z(x)\right)^2 \quad , \qquad y(0) = 1.0$$
$$z'(x) = y(x) - z(x)(1 + z(x)) \quad , \qquad z(0) = 1.0$$

$which\ has\ exact\ solution\quad y(x) = e^{-2x}\ , z(x) = e^{-x}, on\ the\ interval\ [0,1].$ The MAE, ERR presented in Table 3 and Table 4.

7. Conclusion

In this paper, we have described a new algorithm for solving initial value problems of order one, in ordinary differential equations. The implementation of the method is simple and requires two functions evaluation at each iteration. A comprehensive analysis of convergence of the method shows that the order of convergence is quadratic. A stability analysis shows that method is A-stable. The computational results obtained by our method compares favorably with other method. Our future works will deal with extension of the present method to solve higher order boundary value problems and to improve order of accuracy.

Table 1.

N	MAE		ERR	
	Method(2.7)	Method[7]	Method(2.7)	Method[7]
512	.19214574(-2)	.22224213(-2)	.19202513(-2)	.21395404(-2)
1024	.48198143(-3)	.57727523(-3)	.47781889(-3)	.55930193(-3)
2048	.12194760(-3)	.14899758(-3)	.11446897(-3)	.14826830(-3)
4096	.32158459(-4)	.47431273(-4)	.20055209(-4)	.46225156(-4)
8192	.92562505(-5)	.12425815(-4)	.81342805(-5)	.85550200(-5)
16384	.31134662(-5)	.63391294(-5)	.24122351(-5)	.14024622(-5)

Table 2 .

N	MAE		ERR	
	Method(2.7)	Method[7]	Method(2.7)	Method[7]
16	.67436675(-4)	.74768046(-4)	.67436675(-4)	.74768046(-4)
32	.16772730(-4)	.18680079(-4)	.16772730(-4)	.18680079(-4)
64	.41365452(-5)	.43749637(-5)	.41365452(-5)	.43749637(-5)
128	.61987117(-6)	.85828975(-6)	.61987117(-6)	.85828975(-6)

Table 3.

N	MAE(Y)		MAE (Z)	
	Method(2.7)	Method[7]	Method(2.7)	Method[7]
64	.30219555(-4)	.74401498(-4)	.64671040(-5)	.12636185(-4)
128	.98347664(-5)	.16763806(-4)	.12516975(-5)	.32186508(-5)
256	.25033951(-5)	.52154064(-5)	.80466270(-6)	.83446506(-6)
512	.15497208(-5)	.11771917(-5)	.74505806(-6)	.41723251(-6)
1024	.71525574(-6)	.12069941(-5)	.47683716(-6)	.47683716(-6)
2048	.65565109(-6)	.77486038(-6)	.10132790(-5)	.10132790(-5)

Table 4.

N	ERR(Y)		ERR (Z)	
	Method(2.7)	Method[7]	Method(2.7)	Method[7]
64	.30219555(-4)	.74401498(-4)	.64671040(-5)	.12636185(-4)
128	.94026327(-5)	.16763806(-4)	.12516975(-5)	.32186508(-5)
256	.12218952(-5)	.51707029(-5)	.74505806(-6)	.83446503(-6)
512	.84936619(-6)	.11026859(-5)	.62584877(-6)	.26822090(-6)
1024	.64074993(-6)	.11771917(-5)	.23841858(-6)	.17881393(-6)
2048	.29802322(-7)	.17881393(-6)	.77486038(-6)	.77486038(-6)

References

[1] Ahmad, R. R,Yacoob, N. and Mohd. Murid, A. H.: *Explicit method in solving stiff ordinary differential equations*, Int. J. Comput. Math. 81(2004), pp. 1407-1415.

[2] Conte, S. D. and de Boor, C.: *Elementary Numerical Analysis, An Algorithmic Approach,* McGraw Hill, New York (1980).

[3] Jain, M. K., Iyenger, S. R. K. and Jain, R. K.: Numerical Methods for Scientific and Engineering Computation (2/e), Wiley Eastern Ltd. New Delhi (1987).

[4] Lambert, J. D.: *Numerical Methods for Ordinary Differential Systems*, John Wiley, England (1991).

[5] Ramos, H.: *A non - standard explicit integration scheme for initial value problems*, Applied Mathematics and Computation, 189 (2007), pp.710-718.

[6] Sunday, J. and Odekunle, M. R. : *A New Numerical Integrator for the Solution of Initial Value Problems in Ordinary Differential Equations, The Pacific Journal of Science and Technology,* Vol. 13 ,No. 1 (2012),pp.221-227.

[7] Van Niekerk, F. D.: *Nonlinear one step methods for initial value problems,* Comput. Math. Appl. 13 (1987), pp. 367-371.

[8] Wu, X. and Xia, J.: *Two low accuracy methods for stiff system*, Appl. Math. Comput., 123 (2001), pp.141-153.

* Present Address: Department of Information Technology, College of Applied Sciences P.Box. 1905, PC 211, Salalah, Sultanate of Oman.