Acta
Technica
Jaurinensis

Vol. 9, No.1, pp. 1-15, 2016
DOI: 10.14513/actatechjaur.v9.n1.383
Available online at acta.sze.hu

Earthquake Risk Assessment – Effect of a Seismic Event in a Moderate Seismic Area

O. Kegyes-Brassai[1], R. P. Ray[1]

[1]Széchenyi István University, Dep. of Structural and Geotechnical Engineering,
Egyetem tér 1., 9026, Hungary
E-mail: kegyesbo@sze.hu, ray@sze.hu

Abstract: This paper presents the process of earthquake risk analysis from the probabilistic determination of seismic hazard and local site effects, through the evaluation of building vulnerability to an event resulting in seismic risk maps. These results can then serve as useful tools for decision makers and insurance companies, and can be applied directly to overall risk management plans of cities. Finally a case study for seismic risk assessment performed in city of Győr is presented, taking into account local site effects based on response analysis with more than 6000 realizations and rapid visual screening of 5000 building to obtain the seismic risk.

Keywords: earthquake hazard, local site effect, vulnerability, risk assessment

1. Introduction: When the Earthquake Comes to Hungary

Hungary is located in an interesting and active geological setting. The Alps and Carpathian mountain ranges give us winter skiing within a short distance. Thermal baths throughout the country are world-famous. The Danube and Tisa rivers make for pleasant travel and recreation sometimes, and are less pleasant other times. The Bakony Region and Lake Balaton are also unique places where one might enjoy the wonders of nature. Mixed into this geological melange is a moderate level of seismic activity. From a seismologist's view, it is not a question of "will there be and earthquake?" but rather "when will the next earthquake come?" Given the level of probability, it is the task of the civil engineer to devise ways to limit the harmful impact of such earthquakes on people. Generally this is what civil engineers do: try to make sense of the unpredictable and dangerous whims of nature while designing many of the building blocks of civilization. With earthquakes and seismic design, the separation between professional practice and public experience is great indeed. In Hungary, there is little, if any "collective memory" of earthquake hazards so the planning and execution of better design and construction practice relies on the efforts of engineers. Other hazards in Hungary, both natural (flooding) and man-made (Kolontár) are still within the public awareness. Earthquakes from 1810, seem like no more of a present threat than Napoleon's army. However, by a systematic method of quantifying the level of seismic hazard in a particular area, one may begin to assess the vulnerability of the surrounding built environment and determine the level of risk the general public faces with respect to economic damage, injury and death. Having a systematic method in place allows decision-makers (government agencies and the public) to better plan, prepare, and devote finite resources to reducing that risk.

2. Earthquakes and Earthquake Damage

Earthquakes and the damage they cause have been recorded throughout history. Ancient civilizations have recorded their occurrence and the havoc they wreak from China to Mesopotamia, to Egypt and Greece. The Carpathian region has witnessed its share of earthquakes as well; over 23000 recordable earthquakes have been found in historical records [22]. Most have been less spectacular events, causing misery to only an unfortunate few. However, what we do know about them is that they have been strong enough and frequent enough to take purposeful action in the design and construction of buildings, bridges, dams, levees, road and railroad embankments. Much of this action is reflected in modern building codes such as Eurocode, where an entire volume is devoted to special design considerations for seismic conditions.

In Hungary, the level of potential seismic shaking is moderate. Using the professional standard for seismic shaking, moment magnitudes can reach Mw 6.2. While this is not nearly as impressive as Mw 9.0 events that have occurred in Japan, Chile, Indonesia and Alaska, they are cause for concern. Examples of moderate events taking place in less-prepared built environments have resulted in genuine disasters, even the Northridge earthquake (1994) with a Mw 6.7 produced a Modified Mercalli Scale IX damage level and was the second costliest disaster in US [17, 19]. In each case, design and construction methods were not adequate for the anticipated earthquake. While perhaps not all the damage could have been prevented, a very large percentage should have gone undamaged. Learning from these events is a very valuable tool for the civil engineer. It presents in very clear terms what sorts of design and construction work and those that do not. This method of back-analysis has led to a great deal of cumulative expertise in the seismic design of structures.

3. A Comprehensive Approach

Seismic design of buildings, bridges, and other structures has evolved to the point where the engineer will take a comprehensive approach to the problem [3]. Defining the magnitude and location of possible seismic events, the general geology of the region, specific land and soil features, the general form of a structure, the materials from which it's made, even architectural details are considered during design. The approach for determining risk is divided into two parts: The seismic hazard (nature at its worst) and building resilience (the constructed environment). The interaction of these two parts may be complex (fig. 1.), but one can imagine that risk is greater when the hazard is strong and the resilience is weak. Since the magnitude and location of a seismic event varies greatly, probabilistic approaches are often used. Previously, engineers would estimate average seismic forces, and produce designs to withstand those forces, with an additional factor of safety added. This deterministic approach works if the designer has a very well-developed intuition about expected seismic forces, past experiments with similar events and structural behavior during earthquakes. A probabilistic approach attempts to quantify certainty (or uncertainty) in the entire design process. Designers can then estimate the level of safety in buildings as well as an estimate of how certain they are and have an understanding how sure they are in that estimate. The general public is less enthusiastic about such an approach. They would rather have reassurances and certainties where none exist. Communicating risk to the public then becomes very important, otherwise, misunderstandings can lead to bad feelings by all as witnessed before and after L'Aquila.

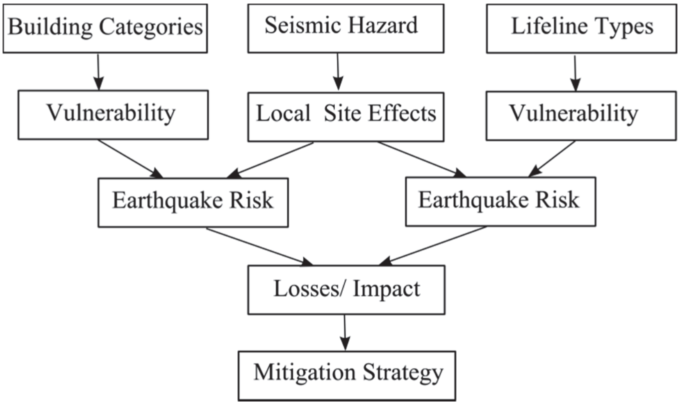


Figure 1. Seismic risk assessment for buildings (left side) and lifelines (right side) [15]

4. Steps to Determine Earthquake Risk

When designing for earthquakes, engineers must first consider how strong that earthquake might be. If we lived in Japan, that earthquake would be intense enough to throw our books off the shelf and destroy our house, if our home was in London it would hardly be enough to rattle our teacups. Hungary lies somewhere in between; in a moderate earthquake zone. In order to quantify the earthquake hazard, the seismic event is divided into segments of its journey to our "home" or building we are designing (fig. 2.). The main components are:

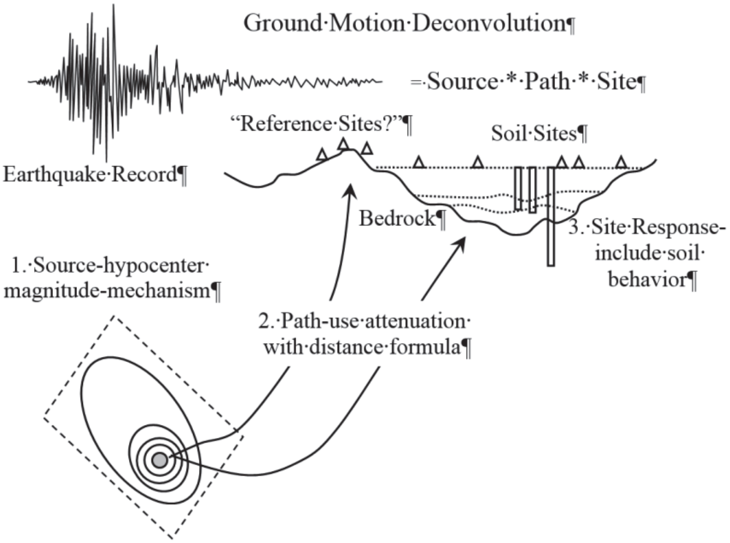


Figure 2. Path of earthquake shaking for seismic hazard assessment (adapted from Boore) [1]

1. Source location and magnitude, determined by the rupture itself.
2. Travel and attenuation through the rock which depends on the properties and layering of rock.
3. Travel and attenuation/amplification through the soils near our structure; the earthquake waves may be trapped, filtered, or amplified by the near-surface (upper 100 m) soils.

4.1. Earthquake Hazards

Seismologists and engineers have a difficult time modeling the very specific source mechanisms. However, they can construct models that faithfully reproduce the location, energy release, and general wave characteristics of the source. Travel through rock layers is based on thousands of seismic records from around the world, and specifically the region of interest.

As they collect more data from different events, seismologists can develop better models to predict the magnitudes and frequencies of earthquake waves that will reach our site. Since local conditions have the greatest effect on our site, they must be considered with great care on a case by case basis. The lithology of local rock and soil above it can magnify earthquake waves many times over or transform them so that there is a predominant frequency of waves that arrive at the site. Naturally, if this frequency is just right (or just wrong) it may prove to be very dangerous to the stability of our structure.

Depending on the degree of refinement we require, we may use a variety of methods to determine the level of hazard at our site. Listing the possibilities from the simplest to most complex:

1. Use an acceleration coefficient (peak ground acceleration, PGA) obtained from

seismic hazard maps and apply it to our static design as an added horizontal load.

2. Use a spectrum of coefficients than vary with frequency and apply the values the most closely match the fundamental frequency of the building.
3. Perform a site response analysis that uses an earthquake record applied to the bedrock in our site and scaled to the proper magnitude (PGA) to see what will happen at the surface.
4. Perform many site response analyses, with an entire family of earthquake records (10-100) and take the average surface response.
5. Determine the statistical distribution of magnitudes (1000's), locations and probability of occurrence of the source earthquake around your region. For each location determine the degree of attenuation of shaking when it reaches your site. Perform site response analysis using every one of these combinations for base input. Determine the statistical distribution of all site responses.

Option 5 is the basis for a probabilistic seismic hazard analysis PSHA. It is the standard analysis method used for critical sites [18]. The PSHA concept is shown in figure 3. Each source location has its own magnitude/recurrence relationship and depending on the geometry of the source location (fig. 3. shows point, line, and area geometries) there will be a spatial probability distribution as well. The distance between the source and the site will influence the level of attenuation of shaking in the rock underlying the site. Combining magnitude/recurrence probabilities and location/attenuation probabilities yields a broad distribution of probability vs magnitude of shaking at the site.

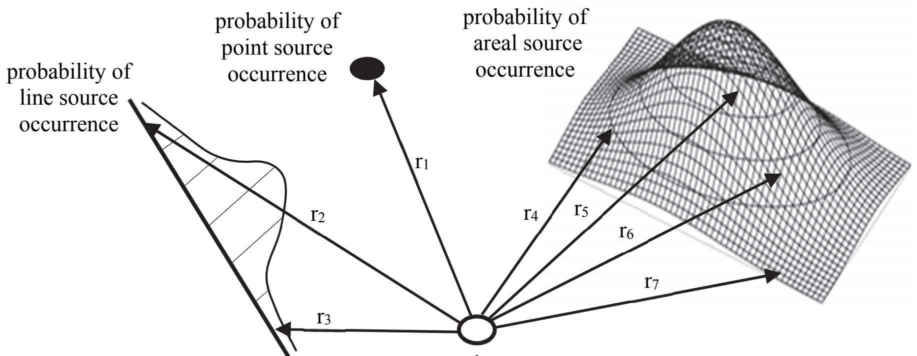


Figure 3. The location of an earthquake may be designated at a point, along a fault line, or over an area. The sum of each location's distribution is P=1.0. Different possible locations of the source mean different attenuation distances.

Performing a PSHA requires a large quantity of data, computing power, good attenuation models (step 2, fig. 2. and fig. 3.), and accurate soil properties for near-surface site response analysis (step 3, fig. 2.). In evaluating seismic response, the engineer is not limited to a single attenuation model or site response method. One may choose from several models and decide which one should be considered the best, then weigh each model with a percentage that totals 100 (e.g. 50%, 35%, and 15%). In this way, we may consider input from a variety of methods to evaluate the same hazard. Using a logic tree approach, one may use different methods along every step of the way.

At the end of this effort, the design engineer has a quantitative measure of the earthquake hazard expressed as peak acceleration, velocity, displacement or as a spectrum of accelerations, velocities and displacements over a range of frequencies. The spectrum approach is useful since it delivers more information than a single peak value, and it can be compared to building response more directly. The spectra may be used as a means of comparing different candidate earthquakes for more rigorous time history building response analysis discussed later. Additionally, the engineer has some quantification of certainty in the calculations. This is represented by the "spread" of the probability distribution as the computations proceed from source to site. The engineer will also have a better idea where more field, laboratory, or analytical work would do the most good in reducing uncertainty.

4.2. Building Resilience

The other side of risk assessment is to determine how resilient the building stock of the examined city may be. Resilience is a combination of strength, stiffness, and ductility which contribute to the performance of a building during an earthquake. One of these quantities alone is not sufficient: a strong building that is not ductile may crack and collapse; a ductile building may sway so much that everything within the structure is destroyed. So the correct combination of these properties, combined with common sense in structural design, will result in very satisfactory performance under seismic loading with only a moderate increase in construction costs. The main component of seismic loading is horizontal. Since buildings are designed to resist vertical forces from gravity, and to a lesser extent wind and ancillary loads applied laterally, the main weakness of most structures is lateral. Horizontal shaking is also stronger than the vertical component. Using an approach similar to evaluating hazards, an engineer may choose from an increasingly complex list of candidates [2]:

1. Apply additional lateral static loading to the structure and compute stresses and deformations to see if they are below maximum criteria.
2. Use additional lateral static load values based on earthquake spectra and the fundamental frequency of the building.
3. Perform a modal analysis to determine the structure's resonant response and apply a better static coefficient for scaling earthquake loads to compute and compare stresses and deflections.
4. Apply an earthquake acceleration history to the base of the structure that represents an expected event and compute stresses and deflections. The structure is modeled as elastic and unyielding. This may sometimes yield unrealistic (too high) stresses on the structure.
5. Push the structure laterally at specific locations until it collapses. Since the analysis is deflection-controlled rather than load-controlled, the structure may soften as it yields. Plastic hinging may take place at several locations and the structure will resist by plastic resilience. This approach, pushover analysis, has become very popular in research and design.

6. Apply an earthquake history to the base of the structure and determine the nonlinear behavior of the entire structure. Such analyses are preceded by linear methods (4) and require placement of plastic hinges at important locations. This is a computationally intense exercise.
7. Perform an analysis similar to (6) but add the nonlinear behavior of the soil and foundation as well. Of course, this becomes extremely time-consuming.

Performing analyses nr. 5, 6, and 7 require multi-processor workstations or high performance servers. A comprehensive pushover analysis (5) may take 1-2 hours for one realization, while (6) would require several hours. A full SSI analysis with nonlinear materials and large deformations would require an overnight computation. High performance computing will significantly reduce the computation time, but still requires hours of careful evaluation per scenario to insure the results are reasonable and properly represent the problem being studied.

The final result of a building analysis may be a simple failure/no failure decision, a capacity curve which can be matched to a hazard curve, or a range of outcomes and probabilities for a single structure or an entire population of structures. Once the building resilience is determined, it can be compared to the hazard. If the hazard is stronger than resilience, failure is possible; if the building is more resilient, it will remain standing. Of course, this is a simplification. Even with the most basic approaches of using a seismic coefficient to compute added static loading, there are elements of probability that are implied, if not stated.

Most work today is based on extensive probabilistic studies, but simplified for direct use through a prescribed evaluation method. The evaluation may be a listing of building features including sized, shape, regularity, and construction materials. Based on different selections, an estimated building resilience is formulated. The advantage of such a method is speed and low cost for applications to a large population of buildings in a city or town. The disadvantage is that other potentially important building features are ignored and the evaluation method is visual so structural details such as reinforcement schedules and joint details are not known. Bear in mind as well that a numerical model, no matter how sophisticated, will not completely capture the behavior of an actual building as it experiences shaking from an earthquake. While precise behavior is very difficult to predict, general behavior and points of strength and weakness throughout the structure do become clearer. Numerical modeling is still more valuable from insight than precise prediction.

4.3. Overall Seismic Risk

The final comparison between hazard and resilience is risk assessment. This not only involves some direct comparison between the level of hazard, i.e. the strength of an anticipated earthquake and structural resilience of the building stock, but also the final outcome of damages and lives lost. Estimating mortality is perhaps the most difficult since there can be so many factors to consider and often little data on which to base estimates. Some of the factors considered in risk assessment are shown in figure 4. such as disaster preparedness, built environment, transportation and communication infrastructure, lifelines, medical readiness, emergency management, economic and social standing, cultural traditions, community and familial ties, construction markets, insurance methods, and many other factors.

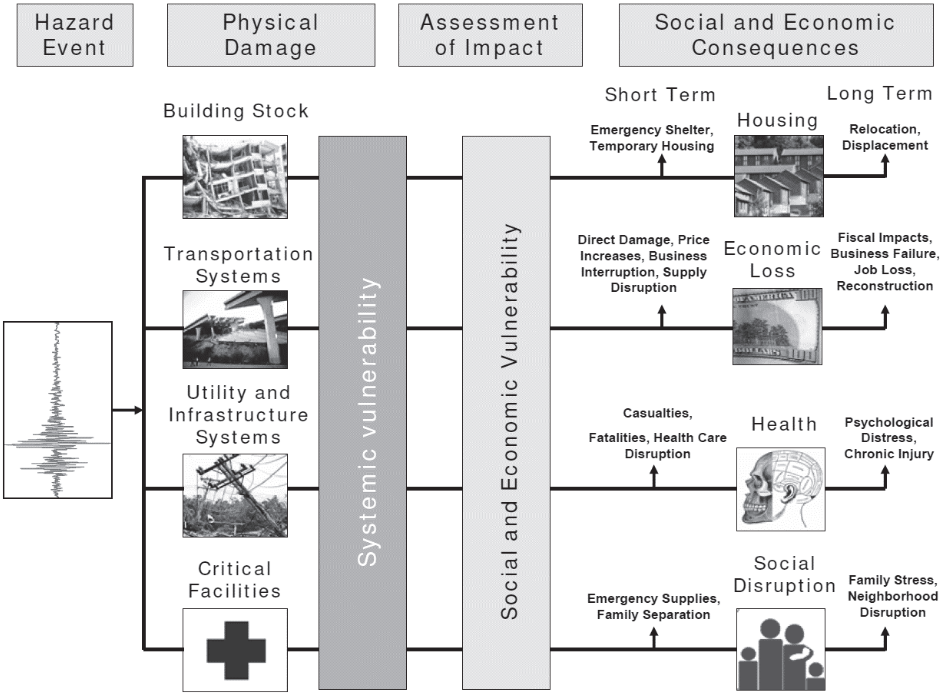


Figure 4. Factors considered in a comprehensive risk assessment [20]

5. Seismic Risk in Hungary

In a region of moderate seismic activity, it would be a prudent exercise to determine seismic risk just as one would consider other natural and man-made hazards. While the specific scientific and engineering knowledge is unique to earthquake phenomena, the application of risk assessment and decision-making is not. Incorporating seismic considerations into the building code has already been done. It has forced engineers to consider the possibility of seismic events that they did not have to consider previously. Of course this is not unique to Hungary, and one of the future challenges is to harmonize evaluation of seismic hazards with neighboring countries in order to have a seamless transition across borders.

5.1. Risk Studies in Hungary

Seismic hazard assessment dates back to 1960s motivated by industrial and economic development requirements. The first probabilistic seismic hazard map was prepared by Zsíros in 1985 [22]. The basic source for earthquake seismic hazard assessment is the Hungarian Earthquake Catalogue [22]. The current Seismic Hazard Map of Hungary [7] is presented in figure 5. It represents estimated contours of maximum horizontal acceleration in bedrock due to seismic events that have a 10% probability of occurring in 50 years. This probability of occurrence is a standard value used throughout the world and means one event with the given value taking into account a return period of 475 years.

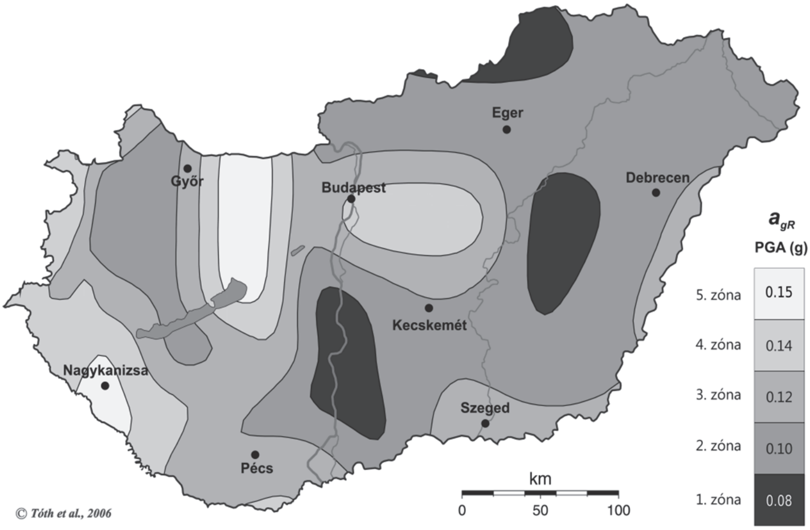


Figure 5. Peak horizontal accelerations from a seismic event that would occur once in 475 years (or a 10% probability of occurring in 50 years). Values range from 0.08g to 0.15g. The highest values are located in the north-south line of the Mór trench and the southwest corner around Nagykanizsa. (adapted from Tóth et al, 2006) [7]

Vulnerability assessments were performed in the case of the nuclear power plant in Paks by T. Katona [12]. Other vulnerability assessment studies have been performed concerning typical masonry and RC buildings in the past few years [13, 14].

Risk assessment studies were performed on the downtown Budapest area [8] and Debrecen [9]. It should be noted that these researches were mainly focused on a deterministic seismic hazard assessment combined with an eigenfrequency estimate of the building stock within 0.1 km2 zones throughout the city. The risk map developed for Debrecen represents a relative degree of damage of the buildings compared to each other.

5.2. Application of Seismic Risk Analysis to Győr

The Győr area is considered to be in a moderate earthquake risk zone with past events estimated up to Mm 6.5. It is a typical situation for many cities in Hungary and throughout Europe where the seismic hazard is not great, but cannot be ignored. Based on the [7] Seismic Hazard Map of Hungary the level of ground motion is 0.12 g PGA at bedrock with 90% probability of non-exceedance in 50 years (fig. 5.). Magnitude 3 and 4 earthquakes are quite frequent in this region, while magnitude 4 and 5 earthquakes happen every 50 years close to Győr, and every 10 years within 100 km of the city. The return period of magnitude 5 and 6 events is about 100 years; this correlates with hazards determined by Tóth et al. [21] and underlines that the seismicity of the region has been underestimated previously by the County Emergency Directorate.

In order to make the best use of limited resources, this methodology used existing soil data, rapid visual building assessment, a limited number of field tests and free, but sophisticated software [15]. As the area of Győr is at the meeting point of different geological structures, further research concerning the local seismicity and soil conditions were performed in order to differentiate between the very diverse territories of the city.

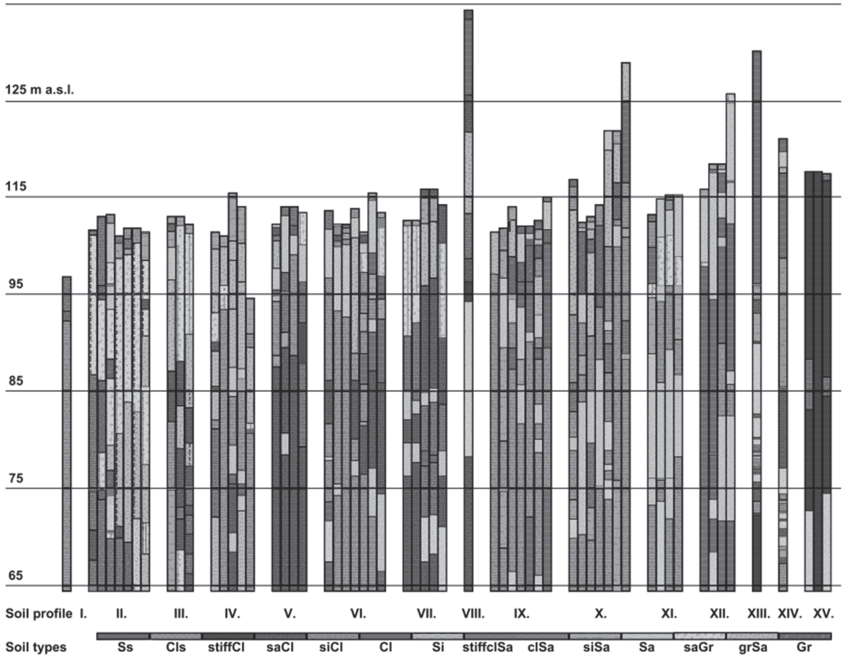


Figure 6. Soil profile groups used to perform detailed site response analyses. Some groups need more data to give a complete picture of that area. [15]

With the permission of the North Transdanubian Environmental Protection and Water Management Inspectorate, soil profiles were collected from the hydrogeological registers. On the territory of the investigated area, around 100 borings were available, from that 60 were picked for further study. The principle of selection was to cover adequately the study area with borings deeper than 30 m. General stratigraphy allowed us to categorize the soils into 15 general profiles as shown in figure 6. In addition to the boring data, field testing by MASW was performed in strategic locations and a limited number of dynamic CPT tests were also applied. Based on the results and correlations with the historical data, soil profile zones were delineated through the city.

Using the different soil profiles, 1-dimensional site response analyses (STRATA [16]) were performed by applying a suite of 7-10 different earthquake records to the base of the profiles. The records were chosen and scaled to fit the general requirements of this region. Additionally, nonlinear behavior models were applied to the different soil layers, based on our dynamic laboratory test results and correlations to other models. The software allowed us to consider reasonable variations in soil properties and thicknesses of layers. In all, 6000 realizations were generated.

The effect of the different soil zones on surface response were evaluated and the results were compared to the more uniform approach described in Eurocode 8 [4]. While there was very general agreement with EC-8, there was also a great deal of variation in the level of seismic action due to the different soil profiles. The different seismic actions could be mapped in a manner similar to the soil profiles. The responses were different enough to divide the city into seven different zones.

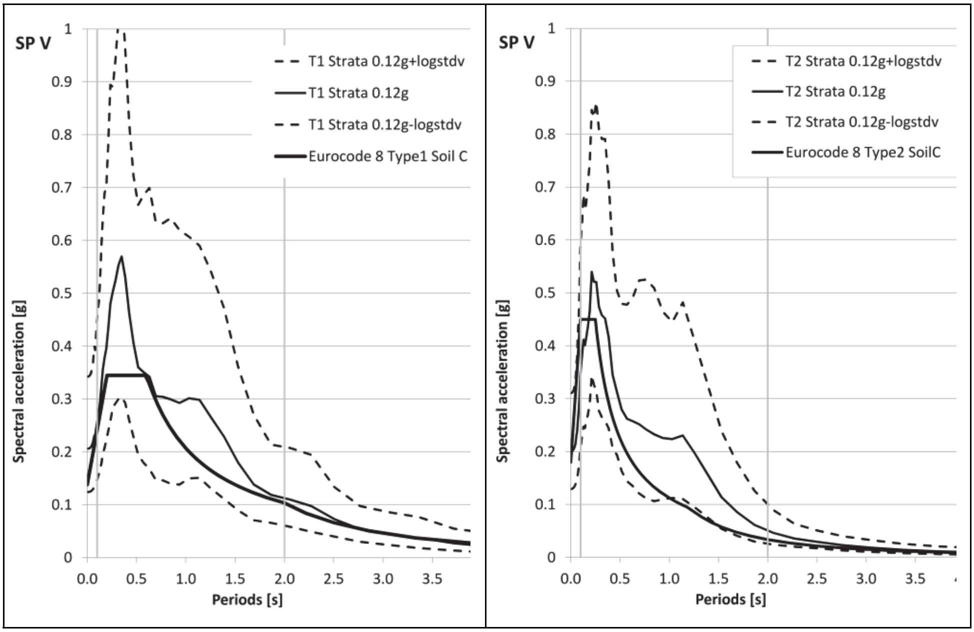


Figure 7. Comparison of average, -1 log standard deviation and +1 log standard deviation results with Eurocode 8 design spectra. Type 1 spectra are shown on the left, Type 2 on the right. The seismic inputs for STRATA matched each Type for comparison.

Soil profile V. is shown in figure 7. where the surface response spectra are shown from the Strata runs. On the left is shown EC-8 Type 1 spectrum for Soil Profile C. Also shown are the average and one log standard deviation spectra from all of the Strata runs performed for this soil profile and using earthquake records that matched Type 1 spectra within a 10-15% range over most of the frequencies. The same is true for the Type 2 spectra comparisons on the right. The PGA for all analyses was 0.12g. While the general magnitude and shape of the surface spectra are similar, there are noticeable differences at the peak values for Type 1 and at slightly lower frequencies (longer period). Each figure represents about 100 realizations from Strata. The use of a log normal distribution to describe variability in response spectra is a standard practice. The figures demonstrate the importance of performing site specific analyses if one needs accurate information. Other soil profiles showed similar variation.

The rapid visual evaluation of buildings was performed on over 5000 structures in Győr. The evaluation method was developed from well-known approaches [5, 6, 10] but modified to account for typical building designs found here. The reliability of the visual evaluations performed by trained non-experts was verified by experts over a significant percentage of all buildings. Further evaluation of building vulnerability was performed by pushover analyses by simplified methods. The pushover analyses showed that the visual vulnerability assessments were consistent and reasonable. The dominant building type found in Győr is load bearing masonry with some reinforced concrete frame components. While they are generally strong buildings, they possess little ductility and do not perform well under severe shaking. When subjected to moderate shaking they will perform adequately much of the time. Construction methods, material quality geometric regularity, roof mass, and presence or absence of reinforced concrete columns and beams all contribute to the buildings' performance.

In order to determine the vulnerability throughout the city, city districts were divided into sections that each had a high percentage of one or two similar building types. This allowed for an accurate and rapid method to divide the risk assessment into manageable pieces. The evaluation process used about 25 sections within the city.

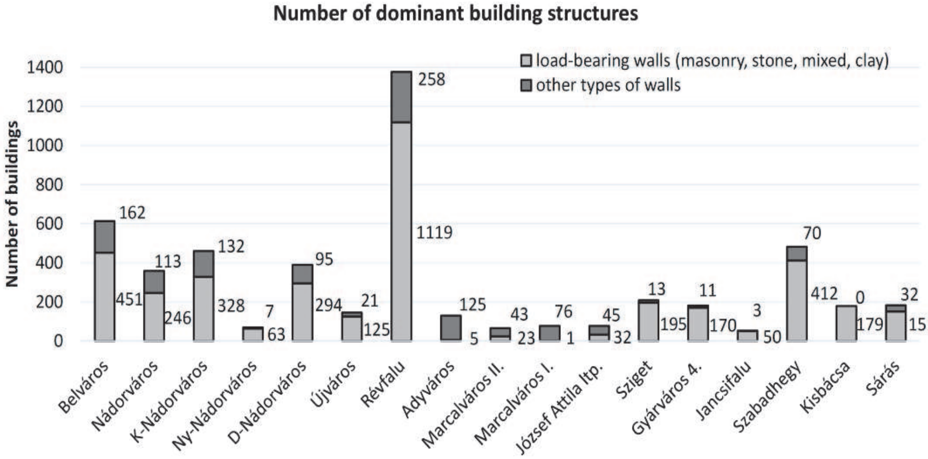


Figure 8. Dominant building types in Győr. Note the very high number of load bearing masonry wall structures.

Estimates of seismic risk were computed using seismic hazard results and building vulnerability scores. As one would expect, since the hazards and vulnerabilities were not uniformly distributed around Győr, there were zones of higher and lower risk. Figure 9. shows the zones that have been delineated. The old downtown area exhibits the highest seismic risk (7) because it has a fairly high hazard level (the site response from its soil profiles were stronger than most) and a high vulnerability (the building stock was generally unreinforced masonry, or combined masonry/concrete load bearing walls). Building geometry was less regular, roofs were generally heavy, and the quality workmanship was highly variable. The industrial park shows the lowest risk (1) because the hazard was moderate; the buildings are modern, framed, reinforced, and regular. Roofs are lighter and workmanship is much more consistent.

Figure 9. Relative seismic risk zones in Győr. Zone 7 is highest risk, Zone 1 is the lowest. The old downtown area has the highest risk because the buildings are old and the relative hazard is high compared to the other areas.

These risks reflect only the structural vulnerability and do not consider the consequences of damage. For example, the destruction of a factory in the industrial district may have a much greater economic impact than the damage done to a downtown building. If this is true, a higher risk would be assigned to the industrial sector. We did not have the necessary economic data and models necessary to perform such an evaluation, but those models do exist [11].

The risk assessment process for Győr is by no means finished. Thorough seismic evaluation of typical building types found in Győr is an ongoing research effort now at Széchenyi University. Further evaluation of buildings, soil properties, anticipated seismic events, and ways to implement positive changes to the design and construction of buildings in Hungary are all needed to improve the quality of life. The study performed here is a small part of a much greater system of hazard evaluation and risk assessment that Civil Engineers do. It is the logical continuation of building engineering that has evolved since the beginnings of civilization. As engineers we are committed to protecting the public safety when it comes to the design and construction of civil works. The process of risk assessment fits perfectly into this commitment and is part of the evolution of techniques and methods that are used to help people make better decisions about what and where to build.

6. Conclusion

In a region of moderate seismic activity, it would be a prudent exercise to determine seismic risk just as one would consider other natural and man-made hazards. While the specific scientific and engineering knowledge is unique to earthquake phenomena, the application of risk assessment and decision-making is not. Incorporating seismic considerations into the building code has already been done. It has forced engineers to consider the possibility of seismic events that they did not have to consider previously. Of course this is not unique to Hungary, and one of the future challenges is to harmonize evaluation of seismic hazards with neighboring countries in order to have a seamless transition across borders.

Earthquake risk is a public safety issue that requires appropriate risk management measures and means to protect citizens, properties, infrastructure and the built cultural heritage. The aim of a seismic risk analysis is the estimation and the hypothetical, quantitative description of the consequences of seismic events of an investigated area, on a regional or state level. The effects to be predicted are the physical damage to buildings and other facilities. Based on damage patterns, further studies can determine the possible number and type of casualties, the potential economic losses due to the direct cost of damage and to indirect economic impacts (loss of the productive capacity and business interruption), the loss of function in lifelines and critical facilities (such as hospitals, fire stations, communication system, transportation networks, water supply, etc.) and also social, organizational and institutional impacts.

Acknowledgement

Data processing of the research concerning seismic risk assessment of Győr was supported by the European Union and the State of Hungary, co-financed by the European Social Fund in the framework of TÁMOP 4.2.4. A/2-11-1-2012-0001 'National Excellence Program' (2013/2014).

References

[1] Boore DM: Can Site Response Be Predicted?, Journal of Earthquake Engineering, Vol. 8, Special Issue 1, pp. 1-41, 2004

[2] Calvi GM et al.: Development of Seismic Vulnerability Assessment Methodologies over the Past 30 Years. ISET Journal of Earthquake Technology, Vol. 43, No. 3, pp. 75-104, 2006

[3] Committee on National Earthquake Resilience: National Earthquake Resilience: Research, Implementation, and Outreach. Washington, D.C.: National Academies Press, 2011

[4] European Committee for Standardization: Eurocode 8: Design of Structures for Earthquake Resistance MSZ EN 1998-1:2004/A1:2013. Brussels: s.n, 2013

[5] FEMA: FEMA 154: Rapid Visual Screening of Buildings for Potential Seismic Hazard. 2002 ed. Washington D.C.: Federal Emergency Management Agency, 1988

[6] FEMA: FEMA 155: SECOND EDITION Rapid Visual Screening of Buildings for Potential Seismic Hazards: Supporting Documentation. 2002 ed. Washington D.C.: Federal Emergency Management Agency, 2002

[7] GeoRisk Earthquake Engineering Ltd.: Seismic Hazard Map of Hungary. Available at: http://georisk.hu/ [Accessed 03 05 2014], 2006

[8] Gribovszki K: Földrengések geofizikai és geológiai környezetének valamint Debrecen földrengésveszélyezettségének vizsgálata térinformatikai eszközökkel (Studying the Geophysical and Geological Environment of Earthquakes and Deterministic Seismic Hazard of Debrecen). Sopron: PhD Dissertation - University of West Hungary, 2005

[9] Gribovszki K, Panza GF: Seismic microzonation with the use of GIS - Case study for Debrecen, Hungary. Acta Geodaetica et Geophysica Hungarica, Vol. 39, No. 4, pp. 177-190, 2004

[10] Grünthal G ed: European Macroseismic Scale 1998. Luxembourg: European Seismological Commission, 1998

[11] Holden RL, Bahls D, Real C: Estimating economic losses in the Bay Area from a magnitude 6.9 earthquake, Monthly Labor Review, December, No.130, pp. 16-22, 2007

[12] Katona T: Seismic Safety Evaluation and Enhancement at the Paks Nuclear Power Plant. Ispra, Italy, Nuclear Energy Agency, Committee on the Safety on Nuclear Installations Workshop Proceedings, 2001

[13] Kegyes-Brassai, O.: Evaluation of Building Structures to Obtain Vulnerability Function in Order to Determine Earthquake Risk. Delft, Fifth International PhD Symposium in Civil Engineering, 2004

[14] Kegyes-Brassai O: The Vulnerability Study of Masonry Structures in Hungary. Geneva, First European Conference on Earthquake Engineering and Seismology joint event of the Thirteenth European Conference on Earthquake Engineering, 2006

[15] Kegyes-Brassai OK: Earthquake Hazard Analysis and Building Vulnerability Assessment to Determine the Seismic Risk of Existing Buildings in an Urban Area, PhD dissertation Széchenyi István University, Győr, Hungary, 199 p, (defense: 2014), 2015

[16] Kottke A, Rathje EM: NEEShub - Resources: STRATA. Available at: https://nees.org/resources/strata [Accessed 10 12 2013], 2013

[17] Martinez M: CNN: Six things we've learned since 1994 Northridge Earthquake. Available at: http://edition.cnn.com/2014/01/16/us/northridge-earthquake-things-learned/ [Accessed 30 04 2014], 2014

[18] Oliveira CS, Campos-Costa A: Overview on Earthquake Hazard Assessment - methods and new trends. In: C. S. Oliveira, A. Roca & X. Goula, eds. Assessing and Managing Earthquake Risk. Dordrecht, The Netherlands: Springer, pp. 15-47, 2006

[19] Southern California Earthquake Center: A Comparison of the February 28, 2001, Nisqually, Washington, and January 17, 1994, Northridge, California Earthquakes. Available at: http://www.scec.org/news/01news/feature010313.html [Accessed 30 04 2014], 2001

[20] SYNER-G: Systemic Seismic Vulnerability and Risk Analysis for Buildings, Lifeline Networks and Infrastructures Safety Gain. 2009-2012. Deliverable 1.1 – SYNER-G work plan Available at: http://www.vce.at/SYNER-G/ [Accessed 25 10 2013], 2009

[21] Tóth L, Győri E, Zsíros T: Seismic Hazard in the Pannonian Region. In: N. Pintér, ed. The Adria Microplate: GPS Geodesy, Tectonics and Hazards. Netherlands: Springer, pp. 369-384, 2006

[22] Zsíros T: A Kárpát-medence szeizmicitása és földrengés veszélyessége: Magyar földrengés katalógus (456-1995) (The Seismicity and Hazard of Carpathian Basin: Hungarian Earthquake Catalogue). Budapest: MTA FKGG Kutatóintézet, 2000

Acta
Technica
Jaurinensis

Vol. 9, No. 1, pp. 16-28, 2016
DOI: 10.14513/actatechjaur.v9.n1.392
Available online at acta.sze.hu

A Method for the Numerical Modelling of Embedded Rails and Determining Parameters to be Optimized

Z. Major

Széchenyi István University, Department of Transport Infrastructure
Egyetem tér 1, 9026 Győr, Hungary
Phone: +36 96 503 038
e-mail: majorz@sze.hu

Abstract: To carry out FEM modelling two model parameters are needed the elastic (Young) modulus (E) and the Poisson-coefficient (μ). Unfortunately, they are not available in every case, and even if so, they need to be corrected. The author's earlier examinations have shown that FEM models are extremely sensitive to the value of the Poisson-coefficient. A very small change in them causes a significant difference in determining the elastic coefficient of the rail support. Literature deals with the factors determining the elastic characteristics of embedded rails. In practice the details for the spring constant on a given geometry test specimen (D) and Shore-A hardness (ShA) are available. These details can't be directly used to create the model but they are suitable to determine parameters that are close to the real values.

Keywords: embedded rail, material properties of pouring material, optimalisation of ERS

1. Introduction

Due to the increasing requirements for rail transport (noise and vibration load, duration costs), embedded rails are gaining more and more ground in our country. They are mainly applied on bridges, grade crossings, tunnels and light rails here. At the moment their practical planning is made with expensive laboratory examinations, examining each variant. Their optimization is genarally not an issue, although accurately selected parameters could lead to significant cost savings and the more punctual performance of the requirements for environmental load.

In this article the author presents the parameters for optimization and their effect on the behaviour of the structure. After that, I am going to introduce the numerical determination of the parameters through two case studies and make a comparison with the results of the real laboratory measurements. Detailed recommendations are found in connection with numerical modelling and laboratory examinations in Stefan Lehner's PhD dissertation [1] and in the author's earlier articles published in Sínek Világa [2, 3]

2. Parameters that can be optimized with FEM modelling

V.L. Markine. A.P. de Man, S. Jovanovic and C. Esveld describe the optimization of embedded rails for multi criteria in their article „Optimum design of embedded rail structure of high-speed lines [4]" and in C. Esveld's book Modern Railway Track [5]. In the article such variants were examined that are unnecessary for practical planning because the variety of available materials and rails is given so their alteration is only possible for determined cases in the engineering practice.(For instance, there is no point in describing the change of thickness of rail foot as a function in the optimization process because these values are characteristic for only one rail system and even with the discovered optimum there will not be a new rail type developed considering the received values.) So the author summarized the characteristics to be optimized below.

- <u>Elastic characteristics</u>: Using properly selected embedding compound, punctually planned embedding sizes, if necessary with the help of elastic trips the bending moment in the rail, the vertical and lateral deformation of the rail under vehicle load, the security against opening gaps because of weld rupture/rail fracture and vertical thermal buckling can be influenced favourably. So during the process of optimization these conditions must be examined together.

- <u>Rail system and rail quality:</u> Due to homogenous, constant vertical and lateral stiffness the bending moment and deformation of the rail are more favourable than in ballasted tracks so it is possible to apply rails with lower inertia systems than in the case of traditional solutions.

- <u>Quantity of embedding compound</u>: the quantity of the embedding compound is not independent from the evolving spring constant, so optimization is only possible the two coefficients are examined together. Quantity is affected by the geometry of the selected rail system, the rail channel, the embedding, and the application of material saving items. Based on the known geometry the rate of the specific material consumption can be determined, which is nothing else but the quantity of the embedding compound referring to one rail running meter. The symbol for the parameter: F_K [dm^3/m].

- <u>Maintenance:</u> On the basis of the maintenance requirements typical for the formed structure, an attempt can be made to determine a parameter describing the maintenance requirements, which takes the interference costs into consideration during the technical lifespan of the rail. It is symbolized with F_F [HUF/m]. In this coefficient the needs can be enforced for rail servicing with different rail types and the favourable effect of the properly selected rail quality.

- Acoustic characteristics: By properly selecting the type of the embedding compound, reducing the free surface of the rail, tearing the vibration bridge more favourable noise and vibration characteristics can be obtained than in the case of traditional tracks. Natural frequency and the free surface are needed to know to make the acoustic qualifying number. There are guidelines for calculation it [4].

With the help of numerical modelling it became possible to model a high number of variations, which can save a significant part of long and expensive laboratory examinations. Using the results of FEM program examinations the optimum variant can be chosen. To achieve this, firstly, it is necessary to normalize each qualifying number, because their order of magnitude and dimension is different, as well. That's why the coefficient of each value is needed to make up with values typical for the initial variant. (A generally applied form that is typical of the given application area can be seen as the initial base, which has to be defined in advance.) After this, by ordering the appropriate sizes of weight to the normalized qualified numbers the variation has to be chosen, in the case of which the total of the weighed qualified numbers is minimal. So the laboratory examinations only have to be carried out on this form if the customer finds it necessary.

3. Data available and necessary for finite element modelling

To carry out FEM modelling two model parameters are needed the elastic (Young) modulus (E) and the Poisson-coefficient (μ). Unfortunately, they are not available in every case, and even if so, they need to be corrected. The author's earlier examinations have shown that FEM models are extremely sensitive to the value of the Poisson-coefficient. A very small change in them causes a significant difference in determining the elastic coefficient of the rail support. Literature [6] deals with the factors determining the elastic characteristics of embedded rails.

In practice the details for the spring constant on a given geometry test specimen (D) and Shore-A hardness (Sh_A) are available. These details can't be directly used to create the model but they are suitable to determine parameters that are close to the real values.

3.1. Determining the elastic module

The value of the elastic modulus can approximately be determined according to the Shore-A hardness. Johannes Kunz and Mario Studer give a closed formula for it [7]. They made observations based on Boussinesq's elastic half space theory matching DIN EN ISO 868 [8], which regulates the geometry of the examination device and the other parameters of the examination. They determined the following relation between the elastic modulus and Shore-A hardness:

$$C = \frac{1-\mu^2}{2 \cdot R \cdot C_3} \cdot \frac{C_1 + C_2 \cdot Sh_A}{100 - Sh_A} \left[\frac{N}{mm^2}\right]. \quad (1)$$

where

E: Elastic modulus of the embedding compound

μ: Poisson-coefficient of the embedding compound

R: radius of durometer

C_1, C_2: examination constants [N]

C_3: examination constant [mm].

The values of the constants are summarized in Table 1. according to the examination standard.

Table 1.: The values of the examination constants according to DIN EN ISO 868

Parameter	***Value***
R	0.395 [mm]
C1	0.549 [N]
C2	0.07516 [N]
C3	0.025 [mm]

The formula above could be used in practice extremely well if it didn't include two unknown coefficients. This problem can be solved with a simplification if μ with 0.5 is substituted. This approach is generally accepted in the engineering practice in case of elastomers. Based on the literature this formula can be applied for hardness values between 30 and 95 Sh_A.

The author applied the method on three embedding compounds available in commercial traffic. He named them A, B and C compounds, making it possible to introduce the method independent from the manufacturer and the product. Making the calculations the received results there are in Table 2. (supposing μ=0.5).

Table 2.: Values of the producing and the calculated elastic modulus

Material	***$Sh_{A\ mean}$***	***$E_{producing}$ [N/mm2]***	***$E_{calculated}$ [N/mm2]***
„A”	40	2.20	2.25
„B”	58	4.20	4.44
„C”	65	5.70	5.90

The results well approach the results given by the manufacturer in spite of the simplification, the difference of the method is only 5.7% even in the worst case.

3.2. Determining the Poisson-coefficient

To make the model produce appropriate outgoing results (vertical stiffness), it is necessary that the Poisson-coefficient should be close to reality. FEM models are extremely sensitive to the changes of the Poisson-coefficient so μ=0,5 approach can't be used, because a stiffer system should be modelled than reasonable. Accepting the elastic module determined earlier, the coefficient can be determined with iteration based on other details.

In the following two chapters the author will only deal with elastic characteristics. The result of the modelling will be compared with the results of two laboratory examinations in order to describe the success of FEM modelling and its usability in the optimization process.

4. Numerical modelling of laboratory examinations

4.1. Examination of structure formed with Ts52 rail

Dr. László Kazinczy announces the examination results of a structure formed with Ts52 rail, VA-70 embedding compound in his presentation [9] in Szeged on 10, April, 2013. The author determined the typical parameters of the embedding compound (E, μ) with the method shown in the previous chapter and based on the stated geometry (Figure 1.). The modelled structure is shown in Figure 2..

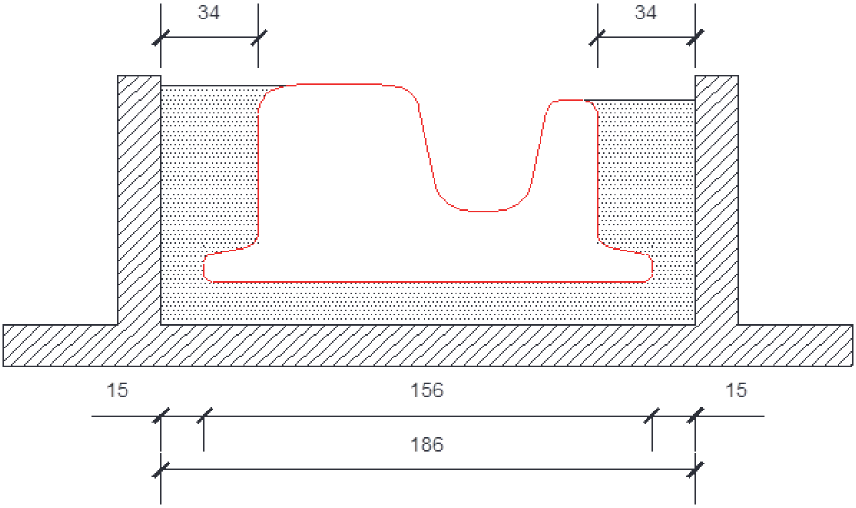


Figure 1.: The geometry published by Dr. László Kazinczy (the sizes are counted in mm)

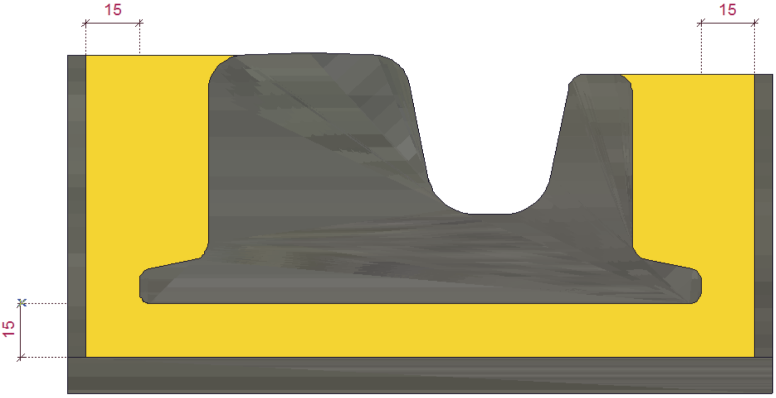


Figure 2.: The modelled rail channel (Ts52) (the sizes are counted in mm)

Two models are needed to prepare for examinations. A 10-mm-thick plain strain model was applied in the condition of planar deformation to determine the static embedding coefficient and a 250-mm-long plate model was used to examine the pushing.

To determine the embedding coefficient the plain strain model with 1 [kN] vertical force was loaded. The results of running are shown in Figure 3..

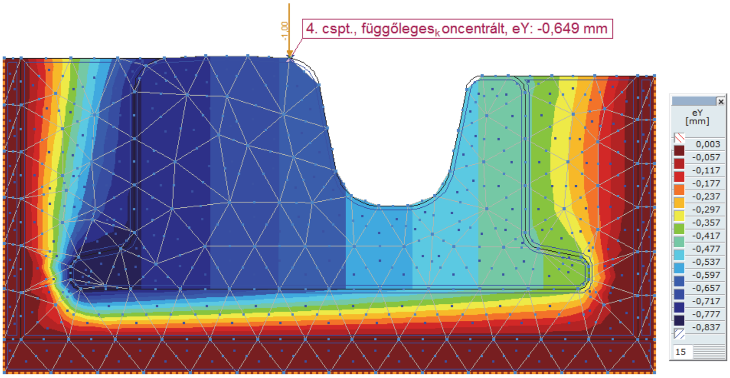


Figure 3.: Vertical deformation in the plain strain model

Based on the received results the embedding coefficient can be calculated with the following formula:

$$C = \frac{1000\,[\mathrm{N}]}{0.649\,[\mathrm{mm}] \times 10\,[\mathrm{mm}] \times 156\,[\mathrm{mm}]} = 0.99 \left[\frac{\mathrm{N}}{\mathrm{mm}^3}\right] \quad (2)$$

The following table summarizes the results that Dr. Kazinczy László and I received:

Table 3.: The embedding coefficient value in the examined cases (Ts52) (I.)

Case	***Embedding coefficient [N/mm3]***
250 mm - Test specimen - Dr. Kazinczy	1.18
900 mm - Test specimen - Dr. Kazinczy	0.91
Mean value- Dr. Kazinczy	1.05 (100.0 %)
FEM model (10 millimeter)	0.99 (94.3 %)

The calculation was repeated in order to get a picture of the behaviour of the model. In these cases 1 [kN] force is placed on the 250 and the 900-mm-long plain strain model. The received results are summarized in Table 4..

Table 4.: The embedding coefficient value in the examined cases (Ts52) (I.)

Case	***Embedding coefficient [N/mm3]***
Mean value - Dr. Kazinczy	1.05 (100.0 %)
FEM model (250 mm)	0.99 (94.3 %)
FEM model (900 mm)	0.89 (84.8 %)

Based on the numerical examinations carried out it can be seen that the size of the model has an effect on the received results. In the 10-250-mm-long scope the difference is 5.7%, whereas this difference is 15.2% in case of 900 mm. It has to be noted that if measured values with the average are compared, the difference is 12.4% in case of the 250-mm trial body, and it is 13.4% with the 900-mm one. The results above show that the behaviour of the applied model is very similar to the behaviour of the real structure.

After determining the embedding coefficient the longitudinal resistance were examined on the 250-mm-long plate model. 96.18 (kN) force was put on the model and the extent of the occurring displacement was examined. The results of the running are described in Figure 4..

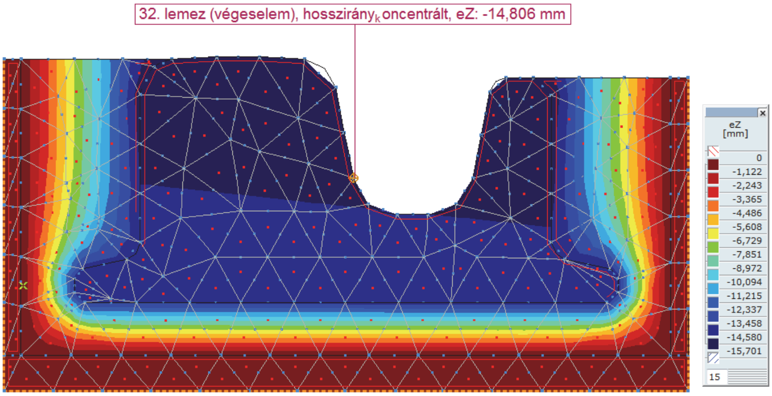


Figure 4.: Longitudinal deformation on the plate model (Ts52)

The results that Dr. László Kazinczy and the author received are summarized in Tables 5. and 6..

Table 5..: The values of longitudinal displacement in the examined cases (Ts52)

Case	***Longitudinal displacement [mm]***
250 mm- Test specimen - Dr. Kazinczy	16.000 (100.0 %)
FEM model	14.806 (92.6 %)

Table 6.: The value of the longitudinal spring constant in the examined cases. (Ts52)

Case	***Longitudinal spring constant [kN/mm]***
250 mm- Test specimen - Dr. Kazinczy	6.01 (100.0 %)
FEM model	6.49 (107.9 %)

Based on the results above it can be seen that the results received by FEM modelling adequately approach the results of real laboratory measurements. (There has not been a bigger difference than 8% in the examined case), so they make the appropriate start-up base for the optimization process. More research is needed to select the right size of the model and to determine the exact behaviour of the embedding compounds with different stresses.

4.2 Examination of a structure with 35GPB rail

Dr. Ferenc Horvát and Dr. György Németh referred to the examination report No R3830 of the Edilon company in their certification [10] published on 26 April 2004. During the examinations they applied a test specimen formed with 35GPB rail and VA-60 N compound.

The width of the rail channel was 170 mm and its height was 97 mm. Under the rail foot the thickness of the pouring was 15 mm. The length of the trial body was 300 mm. The prepared model is shown in the following Figure 5..

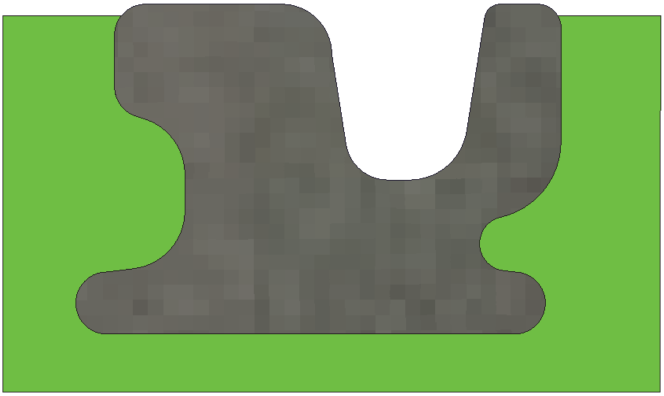


Figure 5.: The modelled rail channel (35GPB)

When examining the vertical static spring constant a 300-mm-thick plain strain condition model was used, which was loaded with 22.5 (kN) vertical force. The displacements of the model are shown in Figure 6..

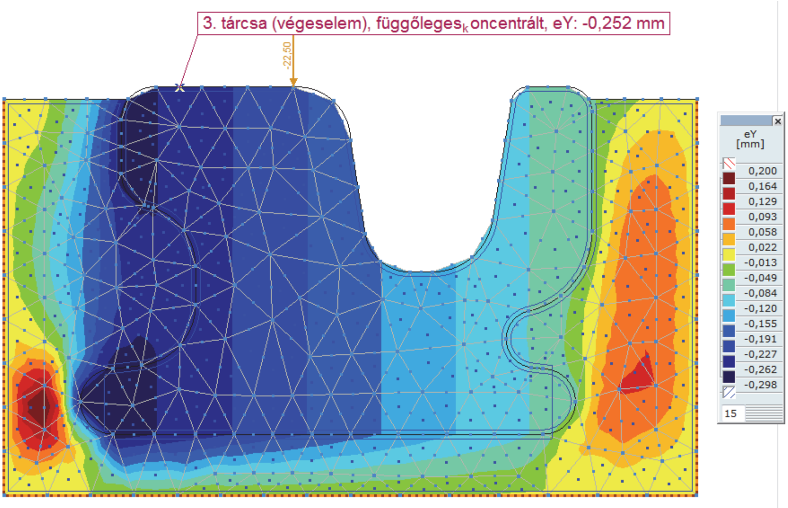


Figure 6: Vertical deformation of the model (35GPB)

The received results are summarized in Table 7..

Table 7.: The value of vertical displacement in the examined cases (35GPB)

Case	***Vertical displacement [mm]***
Report - R3830	0.290 (100.0 %)
FEM model	0.252 (86.9 %)

Based on the results the value of the vertical static spring constant for a 300-mm test specimen was determined. The results received are summarized in Table 8..

$$k_{stat} = \frac{22.5\,[\mathrm{kN}]}{0.290\,[\mathrm{mm}]} = 77.6\,[\mathrm{kN/mm}] \quad (3)$$

$$k_{stat} = \frac{22.5\,[\mathrm{kN}]}{0.252\,[\mathrm{mm}]} = 89.3\,[\mathrm{kN/mm}] \quad (4)$$

Table 8.: The value of the vertical static spring constant in the examined cases (35GPB)

Case	***Static vertical spring constant [kN/mm]***
Report - R3830	77.6 (100.0 %)
FEM model	89.3 (115.1 %)

In the examined case the difference from the measured result was 15.1 %. The examinations with the Ts52 rail have shown that the differences of the 2 measurements from the average are 12.4% and 13.4% in real cases. Based on this it can be stated that the behaviour of the applied model well approaches the behaviour of the real structure.

Finally the extent of resistance against longitudinal force on a 300-mm-long plate model was examined. The occurring displacement is shown in, Figure 7., the results received are summarized in Table 9. and 10..

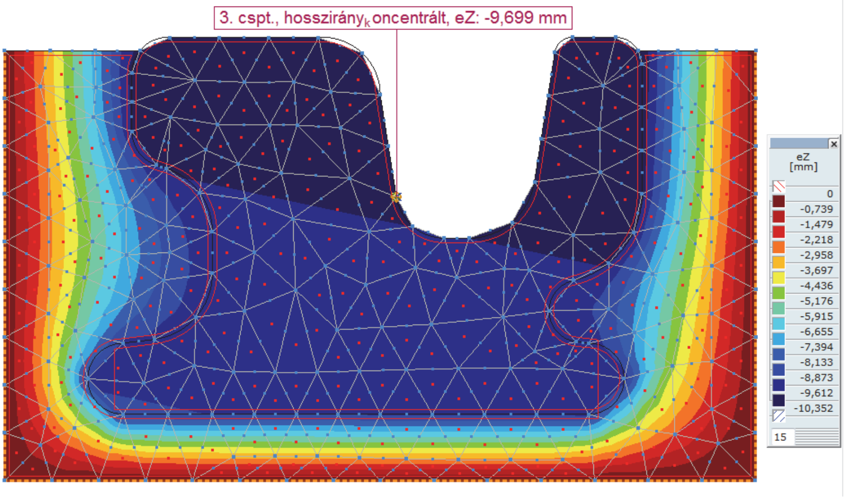


Figure 7.: Longitudinal deformation on the plate model (35GPB)

Table 9.. Longitudinal displacement in the examined cases (35GPB)

Case	***Longitudinal displacement [mm]***
Report - R3830	10.000 (100.0 %)
FEM model	9.699 (96.9 %)

Table 10.: The value of the longitudinal spring constant in the examined cases

Case	***Longitudinal spring constant [kN/mm]***
Report - R3830	5.10 (100.0 %)
FEM model	5.16 (101.2 %)

Based on the results above it can be seen that the results of FEM modelling well approach the results of real laboratory measurements. (The greatest difference was about 15% in the examined case), so they make an adequate initial base for optimization. More research is necessary to select the appropriate size of the model and determine the exact behaviour of the embedding compounds with different stresses.

5. Further research possibilities

FEM modelling shown in this article offers good opportunity for investigation and design of railway superstructure with embedded rails. It is an important question to accurate design of transition zones. In this case FEM modelling is not enough, because spring constants related to ballasted tracks can't be determined. In case of modelling ballasted tracks the DEM simulation is able to be applied [11, 12, 13], in which railway ballast is modelled by discrete element particles (balls and clumps). Layer structures can be optimised by using results that come from DEM models. This is very important because track geometry faults can be avoided. If track faults aren't eliminated speed restrictions have to be used that cause additional traction energy and time consumption of railway vehicles in operation [14, 15]. These energy and time consumption values can be calculated according to the allowed speed and type of railway vehicles, as well as through-rolled tons.

6. Summary

In this article the author presented a method that is suitable for modelling the elastic behaviour of embedded rails with adequate accuracy. To show the success of the modelling the author compared the results of two laboratory examinations with the results of FEM model. It has been proven that the results given by the models provide a good initial base for the optimization of embedded rail structures. Only a part of the necessary details are available for optimization so more research is needed to determine the adequate weight coefficients to find out about optimization factors on the basis of the contractor's and the operator's details.

References

[1] Lehner S: Kontinuierlich eingegossene Schiene, Temperaturverteilung-Verbundwirkung_Brücken

[2] Major Z: Rugalmas ágyazású kiöntött csatornás vasúti felépítmény (1. rész): A síncsatorna kiöntőanyag jellemzőinek meghatározása, Sínek Világa, Vol. 55, No. 6, pp. 10-12, 2013

[3] Kulcsár N, Major Z: Rugalmas ágyazású kiöntött csatornás vasúti felépítmény (2. rész): Rugalmas síncsatorna-kiöntések numerikus modellezése, Sínek Világa, Vol. 56, No. 1, pp. 22-26, 2014

[4] Markine VL, de Man AP, Jovanovic S, Esveld C: Optimum design of embedded rail structure of high-speed lines, CD Proceedings of the International Conference, Railway Engineering 2000, London UK, July 2000

[5] Esveld C: Modern Railway Track – Second Edition, MRT-Productions, Zaltbommel, 2001

[6] Ludvigh E: Elastic behaviour of continuously embedded rail system, Periodica Polytechnica, Vol. 46, No. 1, pp. 103-114, 2002

[7] Kunz J, Studer M: Determining the Modulus of Elasticity in Compression via the Shore A Hardness, Kunststoffe international, No. 6, 2006

[8] DIN EN ISO 868: Kunststoffe und Hartgummi- Bestimmung der Härte mit einem Durometer (Shore- Härte)

[9] Kazinczy L: Sínszálak ágyazásának és rögzítésének körülményei a burkolt közúti vasúti vágányoknál – presentation, Szeged, 10 April 2013.

[10] Horvát F, Németh Gy: Hazai Megfelelőség Igazolás – 35GPB (35LPG) r. sínekkel kialakított, Edilon típusú kiöntött síncsatornás közúti vasúti vágány kísérleti szakaszának építésére, Győr, 26 April 2004.

[11] Fischer Sz, Horvát F: A georács erősítésű vasúti zúzottkő ágyazat diszkrét elemes modellezési lehetőségei, Közlekedésépítési Szemle, Vol. 60, No. 8, pp. 20-29, 2010

[12] Fischer Sz, Horvát F: Investigation of the reinforcement and stabilisation effect of geogrid layers under railway ballast, Slovak Journal of Civil Engineering, Vol. 19, No. 3, pp. 22-30, 2011
DOI: 10.2478/v10189-011-0015-y

[13] Fischer Sz, Horvát F: Superstructure stabilization of ballast bedded railway tracks with geogrids, Hungarian Journal of Industrial Chemistry, Vol. 39, No. 1, pp. 101-106, 2011

[14] Fischer Sz: Traction Energy Consumption of Electric Locomotives and Electric Multiple Units at Speed Restrictions, Acta Technica Jaurinensis, Vol. 8, No. 3, pp. 240-256, 2015
DOI: 10.14513/actatechjaur.v8.n3.384

[15] Fischer Sz: Lassújel miatti többlet-költségek és a megszüntetés költségeinek összehasonlítása, Sínek Világa, No. 5, pp. 21-29, 2011

Acta
Technica
Jaurinensis

Vol. 9, No. 1, pp. 29-41, 2016
DOI: 10.14513/actatechjaur.v9.n1.391
Available online at acta.sze.hu

Special Thermal Insulation Methods of Building Constructions with Nanomaterials

D. Bozsaky

Széchenyi István University, Department of Architecture and Building Construction
Egyetem tér 1, 9024 Győr, Hungary
Phone: +36 96 503 454
e-mail: bozsaky@sze.hu

Abstract: Nanotechnology-based materials have previously been used by space research, pharmaceuticals and electronics, but in the last decade several nanotechnology-based thermal insulation materials have appeared in the building industry. Nowadays they only feature in a narrow range of practice, but they offer many potential applications. These options are unknown for the most architects, who may simply be afraid of these materials owing to the incomplete and often contradictory special literature. Therefore, they are distrustful and prefer to apply the usual and conventional technologies. This article is intended to provide basic information about nanotechnology-based thermal insulation materials for designers. It describes their most important material properties, functional principles, applications and potential usage options in building construction.

Keywords: nanotechnology, thermal insulation, building construction

1. Introduction

The prefix “nano” is derived from the Greek “nanos”, meaning "dwarf". Nano (symbol: n) is an SI prefix meaning one billionth. In the metric system this prefix denotes a factor of 10^{-9}. Nanotechnology is science, engineering, and technology conducted on a nanoscale, which is about 1-100 nanometres. It is the study and application of extremely small things and can be used across other science fields, such as chemistry, biology, physics, materials science, and engineering. Its traditional sense means building things from the bottom up, with atomic precision [1].

2. Brief History of Nanotechnology

Biological systems often feature natural, functional nanomaterials. The structure of foraminifera and viruses, the wax crystals covering a lotus leaf, spider and spider-mite silk, some butterfly wing scales, natural colloids (milk, blood), horny materials (skin, claws, beaks, feathers, horns, hair), corals, and even our own bone matrix are all natural organic nanomaterials [1].

Natural inorganic nanomaterials occur through the growth of crystals in the diverse chemical conditions of the earth's crust. For example, clays display complex nanostructures due to the anisotropy of their underlying crystal structure, and volcanic activity can give rise to opals, which are an instance of naturally occurring photonic crystals due to their nanoscale structure.

The concepts that seeded nanotechnology were first discussed at an American Physical Society meeting at the California Institute of Technology (California, USA) on 29th December, 1959, by the renowned physicist Richard P. Feynman (1918-1988) in his talk *There's Plenty of Room at the Bottom*, in which he described the possibility of synthesis via the direct manipulation of atoms. The term "nanotechnology" was first used by the Japanese scientist, Norio Taniguchi (1912-1999) in 1974, though it was not widely known. The first fundamental studies about nanotechnology were written by Claes-Göran Granqvist (1946-) and Robert A. Buhrman (1944-) in 1976 [2].

However, the term was not used again until 1981 when Kim Eric Drexler (1955-), who was unaware of Taniguchi's prior use of the term, published his first paper on nanotechnology. He popularized the concept of nanotechnology and founded the field of molecular nanotechnology. In his 1986 book, *Engines of Creation: The Coming Era of Nanotechnology*, he proposed the idea of a nanoscale assembler. Also in 1986, Drexler co-founded the Foresight Institute to help increase public awareness and understanding of nanotechnology concepts and implications [2].

However, when something is as small as an atom, it is impossible to see it with the naked eye. In fact, it is impossible to see with the microscopes typically used in high school science classes. In the 1980s two major breakthroughs initiated the growth of nanotechnology in the modern era. The first was the invention of the scanning tunneling microscope (STM) in 1981 by Gerd Binning (1947-) and Heinrich Rohrer (1933-2013) at the IBM Zurich Research Laboratory. It provided an unprecedented visualization of individual atoms and bonds and was successfully used to manipulate individual atoms in 1989. The second invention, fullerenes (Buckminsterfullerene: C_{60}), were discovered in 1985 by Harold Walter Kroto (1939-), Richard Errett Smalley (1943-2005), and Robert Floyd Curl (1933-) at Rice University. C_{60} was not initially described as nanotechnology. The term was used regarding subsequent work with related graphene tubes, which suggested potential applications for nanoscale electronics and devices [1].

3. Nanotechnology in Architecture

Nanotechnology can also be used in architecture. The addition of nano-silica to cement-based materials can increase their durability and compressive strength. It can be used to increase the fluidity or water permeability of concrete [3,4]. With the addition of nanotubes or nanofibers, the tensile and bending strength of concrete structures can be strengthened [5]. Wood can be composed with nanotubes or nanofibers, and these products can be twice as strong as steel [2]. Titanium dioxide (TiO_2) is used in nanoparticle form to coat glazing because of its sterilizing and anti-fouling properties

[4,6]. Nanoparticle-based coatings can provide better adhesion, transparency, self-cleaning, corrosion and fire protection.

We can also find several nanotechnology-based thermal insulation materials on the market:

- Expanded polystyrene products, including graphite powder additive (EPS Graphite)
- Aerogel used for insulating transparent building structures
- Nanoparticle-based vacuum insulation panels
- Nano-ceramic thermal insulation coatings

4. Nanotechnology-Based Thermal Insulation Materials

Nanotechnology-based thermal insulation materials generally have a better thermal insulation quality than traditional materials. There are three ways of heat transport in traditional thermal insulation materials: thermal conduction (vibration of molecules inside cellular walls), heat flux (between air particles enclosed in cells) and thermal radiation (between opposing cellular walls). In nanotechnology-based thermal insulation materials one or more ways of heat transport is hampered and uneasy. This is why they can decrease the heat transfer coefficient of building structures. The well-known formula of heat transfer coefficient is the following:

$$U = \frac{1}{\frac{1}{h_i} + \sum_{i=1}^{n} \frac{d_i}{\lambda_i} + \frac{1}{h_e}} \tag{1}$$

In the formula U is the heat transfer coefficient [W/m^{2}K]; d_i is the individual thickness of each structural layer [m]; λ_i is the individual thermal conductivity of each structural layer [W/mK]; h_i is the internal convective heat transfer coefficient [W/m^{2}K]; and h_e is the external convective heat transfer coefficient [W/m^{2}K].

This U-value can be decreased in two ways by nanotechnology-based materials. Some of them (EPS Graphite, aerogel insulations, vacuum insulation panels) have a lower thermal conductivity than the traditional materials, because the ways of heat transport is blocked inside them. In EPS Graphite products only the thermal radiation is slowed down. In vacuum insulation panels the thermal conduction and heat flux are also balked. In aerogel insulations all the three ways of heat transport are blocked. Other materials (e.g. nano-cermaic thermal insulation coatings) can reduce the convective heat transfer coefficient (h_i, or h_e, according to the insulated side) of the insulated surface.

4.1. Expanded Polystyrene Products, Including Graphite Powder Additive

Expanded polystyrene products including graphite powder additive (hereafter, “EPS Graphite”) are applied to a wide range materials in building construction (e.g. Austrotherm Grafit®). At the same time a foaming agent (pentane) is added, micro-size graphite powder is mixed with polystyrene during production. Besides the grey colour the end product is different from the traditional expanded polystyrene products because of the 20% lower thermal conductivity. In this material one way of heat transport is balked because thermal radiation is partially diffused and reflected by the graphite powder [7].

The other physical parameters of EPS Graphite (Table 1) are the same as traditional EPS products, so the applications are also similar. Their construction needs special attention. Because of their dark colour EPS Graphite can collect much more heat than traditional products. This is why higher thermal movements may occur in them. In the case of strong sunshine it is recommended, that it be placed in shaded surfaces (following the sun). Surface-reinforcing netting should begin as soon as it possible; otherwise, it could become detached from the surface because of the huge thermal movements. A solution to this problem is provided when one side is coated with pink painted EPS Graphite insulating sheets (e.g. Austrotherm Grafit Reflex®). The coating prevents the surface from becoming too warm in bright sunlight. It is important to fasten the boards along their non-painted side on a façade to reflect the sunshine from the surface. Otherwise, we could make a serious design error, because the painted film is unable to enforce its beneficial effect.

Table 1. Main physical parameters of EPS Graphite thermal insulation [7,8]

Physical parameter	***Symbol***	***Dimension***	***Value***
Density	ρ	kg/m^3	15-30
Compressive strength	σ_c	kPa	60-150
Tensile strength	σ_t	kPa	80-100
Bending strength	σ_b	kPa	50-120
Specific heat	c	kJ/kgK	1500
Water absorption	W	V/V %	1-5
Water vapour resistance	μ	-	20-100
Thermal conductivity	λ	W/mK	0.031-0.032
Thermal tolerance	T	°C	85

EPS graphite products are suitable for insulating the inner or outer side of façade walls (inner side with psychometric calculations) and the lower side of cellar slab structures. Products with higher compressive strength (100-150 kPa) are suitable for insulating cellar slab structures (under a concrete underlayment), flat and green roofs with a direct order of layers, and floors lying on the ground (protection against moisture is necessary) [7].

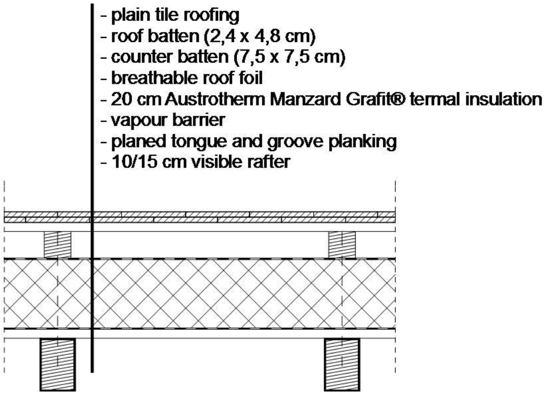


Figure 1. General order of a roof layer, including EPS Graphite (self-made drawing)

Certain products (e.g. Austotherm Manzard Grafit®) are used for insulating high roofs by placing them on the outer side of rafters (Figure 1). Their relatively high compressive strength (150 kPa) allows the structure to withstand the net weight of the roofing and meteorological charges (wind and snow loads). Large cross-sectioned counter battens (75 x 50 or 75 x 75 mm) can be fixed with corrosion-standing fasteners through the insulation. The thermal insulation of high ferro-concrete roofs (coffin slabs) is also possible with this product [7].

4.2. Vacuum Insulation Panels

Vacuum insulation products were first used for insulating refrigerators and vehicles, but nowadays they are used in building construction to provide better insulation performance than conventional materials (Table 2).

Table 2. Main physical parameters of vacuum insulation panels [8]

Physical parameter	***Symbol***	***Dimension***	***Value***
Density	ρ	kg/m^3	150-300
Compressive strength	σ_c	kPa	140-250
Tensile strength	σ_t	kPa	60
Specific heat	c	kJ/kgK	800
Water absorption	W	V/V %	0
Water vapour resistance	μ	-	∞
Liquid water permeability	w	kg/m^{2}h^{0,5}	0
Thermal conductivity	λ	W/mK	0.005-0.010
Thermal tolerance	T	°C	80

In vacuum insulation panels two ways of heat transfer are hampered. The interfaces of silica crystals are very small, which causes difficulties in thermal conduction. Moreover, creating a vacuum (0.05 bar) practically eliminates convection, since this relies on the presence of gas molecules to be able to transfer heat energy by bulk movements. A small decrease in pressure has no effect on the thermal conductivity of a gas, because the reduction in energy-carrying molecules is offset by a reduction in collisions between molecules. However, at a sufficiently low pressure, the distance between the collisions exceeds the size of the vessel, and then the conductivity does reduce with pressure [9].

Vacuum insulation panels consist of a kernel and a membrane. Kernels consisting of compressed fumed silica are laced with silicon carbide (SiC) or other infrared-opacifiers in order to reduce radiative transport and with organic fibres to enhance stability. The kernel is dried, evacuated to about a 0.1 mbar, and sealed into a high-barrier laminate. The laminate consists of several layers of polyethylene-terephthalate (PET) or polyethylene (PE) coated with 30 nm thin aluminium (Al) layers. The Al-layers serve as barriers to air and water vapour. Since such thin metal layers have pinholes, two or three Al-layers are usually integrated into the laminate. Less often, almost water-tight Al-foils with a PE coating as a sealing layer are employed [10].

Depending on the application method, vacuum insulation panels can be installed as normal or mounted panels or as vacuum insulation sandwich panels (VIS). Under certain conditions that use additional protective covers, vacuum insulation panels are suitable for insulating any building construction (flat and high roofs, slab structures, floors, facade walls). In spite of multiple (usually 6-12) layers the protective membrane walls are sensitive, so the full packing of VIP-panels during the production is essential. Even before installation (i.e. when ordering the material), a well-dimensioned product design is necessary because the panels are fully prefabricated. Vacuum insulation panels are delivered to the site in cut-to-size tables, which are designed for the specific surface. Sometimes numbered tables and a laying plan are required [10,11].

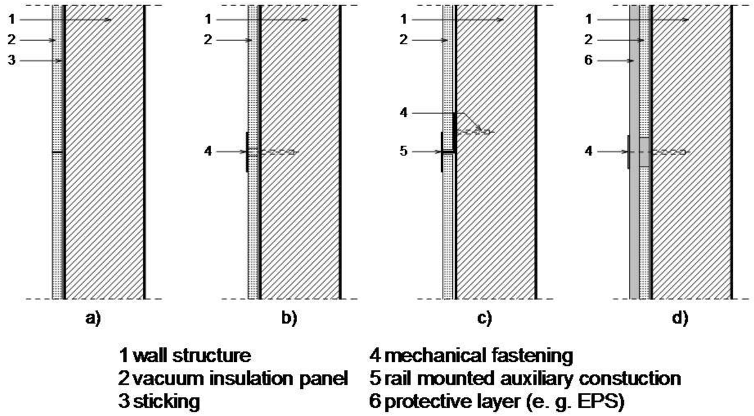


Figure 2. Fixing methods of vacuum insulation panels on façade walls (self-made drawing)

The thickness range of panels is generally 10-50 mm. A panel size from 100 x 100 mm to 1250 x 3000 mm is optional. Their shape should be a rectilinear polygon. Air should not leak into the panels. Therefore it is prohibited to drill through, cut to size, or pierce them subsequently on-site, and the holes of any fixing points are placed in advance. If screw fixation is not necessary (e.g. a floor or flat roof), they can be used without a protective cover. Some kind of fixation is necessary between the plates and the insulated surface on façade walls and the lower side of slab structures. Full surface sticking is possible, but not completely safe, therefore it is generally not recommended. Vacuum insulation panels can be fastened on façade walls with rail-mounted auxiliary construction or wall insulation anchors. Anchors are useable if there are pre-placed holes in the panels for them (Figure 2) [10,11].

In the absence of anchors the use of mounted or prefabricated sandwich panels is recommended because there is no vacuum at the corners of the panels. A recent system is also known as "LockPlate". Its essence is that there is a 4 cm expanded polystyrene foam around the core of a vacuum insulation panel (Figure 3).

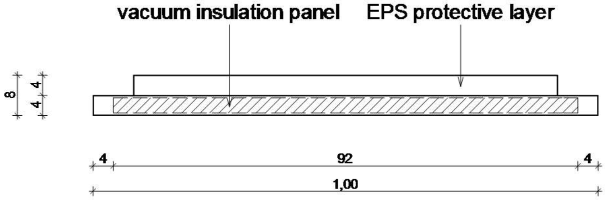


Figure 3. Vacuum insulation "LockPlate" sandwich panel [10]

An adequate quality of the thermal insulation of the parapet parts of curtain wall façades is only available with thick in-fill thermal insulation. When using vacuum insulation panels, much thinner parapet insulation is required, so that the thickness of the parapet and the thermally insulated window can be the same. Vacuum panels can be placed between two layers of glass or metal tables, which change the optical appearance minimally, because the series of transparent windows are only divided by the opaque parapet elements.

4.3. Aerogel Thermal Insulation

Aerogel is a synthetic, porous ultralight material derived from a gel, in which the liquid component of the gel has been replaced with a gas. The result is a solid with extremely low density and low thermal conductivity. It is considered to be the most low-density solid material and has several unique physical properties (e.g. as thermal insulation).

Aerogel was first created by an American chemist, Samuel Stephens Kistler (1900-1975), in 1931, as the result of a bet with Charles Learned over who could replace the liquid in jellies with gas without causing shrinkage. The first aerogels were produced from silica gels. Since then it has been proven that aerogels can be prepared from a number of different materials like alumina, chrome, tin dioxide and carbon [12].

The synthesis of (silica) aerogels may be divided into three general steps: preparation of the gel by sol–gel processes, ageing of the gel in its mother solution to prevent the gel from shrinking during drying, and drying the gel under special conditions to prevent the gel structure from collapsing [13].

Aerogel insulations were developed in the 1980s by Aspen Aerogels, which is a spin-off venture of the National Aeronautics and Space Administration (NASA). The real breakthrough occurred around the year 2000, when reinforcing fibres were successfully into the material successfully. The result was a flexible building material with good technical parameters which is widely applicable in the building industry. Admittedly, it was applied in powder form previously as insulation for large roof windows and other transparent building structures [12].

Silica aerogel is a dendritic network of loose silicon. It is prepared by carefully removing the liquid part from silica alcogel and replacing it with gas, thereby producing the final product containing 99.8% air. It has low thermal conductivity and good physical and mechanical properties (Table 3). It does not age or get mouldy, and environmental problems should not be taken into account during its application. It is hydrophobic (water repellent), water permeable, environmentally safe, and fully recyclable. Its excellent thermal insulation quality is caused by the one-dimensional molecular chain. Air cannot circulate in the pore system of aerogels, so all forms of heat transport are limited.

Thermal conduction slowly occurs inside the nanostructure because of the small interfaces between the aerogel particles. In addition, the pores are smaller than the length of the path required by the internal gas molecules for free collisions. They interfere with cellular walls rather than with each other, so any convective heat transport is minimal. The size of the nanopores is also smaller than the wavelength of infrared rays; therefore, the nanocellular walls reflect and dissolve a significant part of heat rays [4,5].

Table 3. Main physical parameters of aerogel thermal insulation [8,12]

Physical parameter	***Symbol***	***Dimension***	***Value***
Density	ρ	kg/m^3	60-80
Compressive strength	σ_c	kPa	2-100
Tensile strength	σ_t	kPa	200
Specific heat	c	kJ/kgK	750-840
Water absorption	W	m/m %	100
Water vapour resistance	μ	-	5
Thermal conductivity	λ	W/mK	0.013-0.021
Thermal tolerance	T	°C	650

The application fields of aerogel insulations are limited because of the high cost of nanotechnology methods. It is useful for thermodynamic post-improvements of building constructions which have become problematic due to a design or constructional error. It is also utilizable to remedy inherently problematic structural junctions of passive and low-energy houses. It can be applied indoors or outdoors to improve the thermal insulation capacity of monolithic concrete structures (lintels, pillars), window edges, roll covers, pipes, and containers. It is step-resistant and it can usually be installed ina completely ordinary way without disrupting the structure. Stickable, plasterable and flexible, it is easy to fit to curved surfaces. The most important aerogel thermal insulation products used in building construction are:

- Aerogel thermal insulation blankets
- Self-adhesive insulating strips
- Nanoporous translucent glass

An *Aerogel thermal insulation blanket* (e.g. Spaceloft®) is a flexible composite material, which is actually an aerogel embedded in a fiberglass spatial net. It is available in rolls, which enable rapid and wide-ranging applications. Compressible, plastic, and fire-resistant, it fits well anywhere s without loss of tensile strength and elasticity. Aerogel blankets are available in different widths, thicknesses (2-10 mm), and sizes. It can be optimally composed for different temperature ranges (from -200°C to +650°C) [4,5].

The physical parameters of *self-adhesive insulating strips* of aerogel are fully consistent with insulating blankets. They are generally used as thermal break in lightweight building structures (Figure 4). They are also easily and practically applicably and only a 1-2 cm thickness is enough to increase the thermal bridge effect between wall panels and other structural elements [4,5].

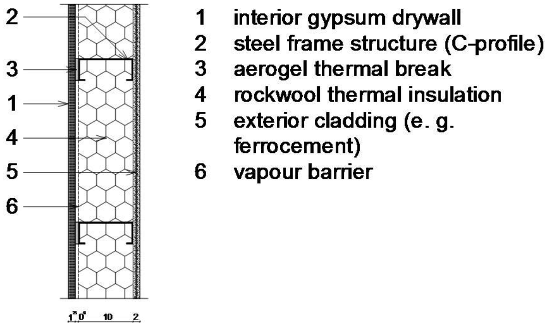


Figure 4. General order of external lightweight wall layers, including an aerogel thermal break [13]

Aerogel is also used as *nanoporous translucent glass* (e.g. NANOGEL®). It is preferably used as a thermal insulating window (e.g. curtain walls), skylight window, transparent partition wall, and balcony railing element [4,5].

Aerogel is especially very interesting as a translucent or transparent insulation material because of its combination of low thermal conductivity and a high transmittance of daylight and solar energy. Two types of highly insulating windows were developed for this purpose based on a granular aerogel and a monolithic aerogel.

Granular aerogel-based windows were developed by ZAE Bayern (Germany). They produced two types of granular aerogel, i.e. semi-transparent and highly translucent granulates. This granular aerogel is stacked in a 16 mm wide polymethyl-methacrylate (PMMA) double-skin sheet between two gaps (i.e. either 12 or 16 mm in width and respectively filled with krypton or argon) and glass panes. Monolithic aerogel-based windows were developed by the HILIT$^+$ project of the European Union. This window is produced in combination with the technology of vacuum glazing by applying a pressure between 1 and 10 mbar [13,14].

4.4. Nanoceramic Thermal Insulation Coatings

In the early 1980s, S. Komarneni and R. Roy developed the first method of synthetizing nanoparticles, specifically nanoceramics [15]. They used a process called “sol-gel” and enabled researchers to test the properties of nanoceramics. This process was later replaced by sintering in the early 2000s and continued to advance to microwave sintering. Because of these advancements, researchers are able to produce nanoceramics at a more efficient rate.

Paint-on insulation products like Thermo-Shield, Protector, MANTI and TSM Ceramics contain microscopic (with a diameter of 20-120 μm) cellular ceramic microspheres. These vacuum-hollow balls were made of high gas-pressure and high temperature (1500 °C) melted ceramic. After they cool down, the pressure ends, leaving a vacuum inside the microspheres. Their binding material is a mixture of synthetic rubber and other polymers. The main components are styrene (20%) and acryl latex (80%). Styrene guarantees the mechanical strength. Acryl latex makes this material resistant against weather conditions and provides adequate flexibility. Other environmental additives (biocides, anti-fouling and antifungal materials) make the final product durable and mould-proofed [16,17,18,19].

Thermal transmission processes happen in these microscopic vacuum spaces in non-traditional ways. There are three ways of heat transfer in thermal insulation materials: thermal conduction (inside cellular walls), heat flux (between air particles enclosed in cells) and thermal radiation (between opposing cellular walls). But there are small interfaces of ceramic microspheres and cellular walls that are so thin they slow down thermal conduction. Heat flux is also unstable in vacuum micro spaces. Air particles come up against cellular walls rather than each other, so they are almost unable to transport heat energy. The inner surface of cellular ceramic microspheres works as a heat mirror and reflects 60-80% of the heat radiation.

The special literature provides different technical details about these materials (Table 4). Moreover, the thermodynamic details are extraordinarily contradictory. Some sources say that their thermal conductivity is around 0.001-0.003 W/mK [16,17,18,19,20], but others publish much higher values (from 0.014 W/mK to 0.140 W/mK) [4,5]. These details are often not confirmed with laboratory tests or refer to inadequate experiments. For example, some sources indirectly determine the thermal conductivity of this thin coating with heat transfer experiments of wall structures according to EN 1934:2000 and EN ISO 8990:2000 standards [16,17,18,19]. Nevertheless, these methods are only suitable for determining the heat transfer coefficient of a global building structure and take account of the standardized heat transmission coefficients on the inner and outer sides of a wall structure. To measure the thermal conductivity of thermal insulation materials, the only suitable standard is EN 12667:2001 (Title: Thermal performance of building materials and products. Determination of thermal resistance by means of guarded hot plate and heat flow meter methods. Products of high and medium thermal resistance) [21]. According to this method, the thermal conductivity was measured in the Laboratory of Building Materials and Building Physics at Széchenyi István University (Győr, Hungary) with a standard heat flow meter of 0.069 W/mK [22]. Based on these experiments, it is concluded that the good thermal insulation quality of nano-ceramic thermal insulation coatings is caused not by very low thermal conductivity but rather by the high convective heat transfer coefficient between the air and surface of the building structure. Further tests are required to determine it exactly [22].

Table 4. Main physical parameters of nano-ceramic thermal insulation coatings [8,16,17,18,19,22]

Physical parameter	***Symbol***	***Dimension***	***Value***
Density (wet)	ρ_{wet}	kg/m^3	500-745
Density (dry)	ρ_{dry}	kg/m^3	290-410
Tensile strength	σ_t	kPa	200-300
Adhesive strength (concrete)	$\tau_{ad,concrete}$	kPa	460-920
Adhesive strength (steel)	$\tau_{ad,steel}$	kPa	470-900
Specific heat	c	kJ/kgK	1080
Water absorption	W	m/m %	20-30
Water vapour resistance	μ	-	2
Liquid water permeability	w	kg/m^{2}h^{0,5}	0.16-0.20
Thermal conductivity	λ	W/mK	0.001-0.003, or 0.014, or 0.070, or 0.140
Thermal tolerance	T	°C	300

Nano-ceramic thermal insulation coatings are flexible, non-toxic, mould-free, UV, fire and chemical resistant, washable, and friendly to the environment; they form a monolithic membrane that bridges hairline cracks. It can stick well to all types of surfaces such as concrete, ceramic, plaster, metal, glass, wood and plastic.

These coatings are typically used for exterior and interior wall insulation, but they are also suitable for pipe insulation and protection against fire and corrosion. They can be easily transmitted to hard-to-reach sites. Their most important advantage is that they can be applied in places where it is not possible to use thick thermal insulating panels (e.g. scheduled national monuments).

After mixing the ceramic microspheres with the binding material, additives and water, a brush, roller or airless spray can be used to apply on the surface to be insulated [16,17,18,19]. To assure adequate and uniform coverage, spray and roll techniques are recommended. Each coat should be sprayed in the same direction to avoid showing undulations and other imperfections on the wall. Very small areas may be brushed. White and any custom colours are available, but darker colours give a correspondingly lower degree of reflectivity [18].

All surfaces must be clean and free from laitance, dust, dirt, rust, oil or grease before painting. Surfaces must be cleaned to remove any loose or chipping paint or other foreign material. No primer layer is usually required, but it is recommended over gypsum board. Generally two layers of nano-ceramic thermal insulation coating are required, the first of which acts as a primer layer. When the coating is applied by brush, three cross-brushed coats are required for adequate insulation. The drying time of a layer depends on the temperature (at 20 °C, it takes 4-5 hours). The complete solidification takes 72 hours [18,19].

5. Conclusions

Nanotechnology-based thermal insulation materials are being used in a wide range of applications in architecture. Due to their non-traditional way of heat transfer ability, the heat flow comes up against several obstructions inside them. This is why they have a higher heat transfer resistance than traditional thermal insulation materials.

EPS graphite products are the most common nanotechnology-based materials. They have a much lower thermal conductivity because of the reduced interior thermal radiation blocked by the fine graphite powder. Formerly, they were only used for façade insulation, and their installation needed much more attention. But due to the product’s improvements, it has overcome its disadvantages and become a widely-used thermal insulation material. The new EPS graphite products with their higher compressive strength are useable for insulation, flat and high roofs, floors and slab structures.

Aerogel insulations and vacuum insulation panels are less widely used thermal insulation materials. They can highly inhibit some forms of heat transfer. Therefore, they have much more heat transfer resistance than traditional materials. Vacuum insulation panels are generally used for insulating building envelopes (façade walls, flat roofs, shutter cases) but only in a narrow range because of their special technological demands. The price of aerogel insulations allows for their application only as a thermal break or insulating glass, but they have many potential applications, with a potential to be widespread in the future.

Nano-ceramic thermal insulation coatings are considered to be the most critical nanotechnology-based thermal insulation material. Architects are quite distrustful of this material. Unlike other materials, nano-ceramic thermal insulation coatings are capable of producing an adequate thermal insulation value of building structures in an extremely thin layer. Their high thermal resistance does not depend on their thickness (like traditional materials) but on their high surface heat transfer resistance. Because of the contradictory technical details in the special literature and the product catalogues of manufacturers, more thermodynamic tests and experiments are needed for these materials, which can provide useful information and technical parameters to practising architects for planning building constructional joints and also utilizable technical parameters for heat transfer and other building energy calculations.

References

[1] Leydecker S: Nano-Materials in Architecture, Interior Architecture and Design, Birkhäuser Verlag AG, Berlin, Germany, ISBN 978-3-7643-7995-7, 2008

[2] McIntyre RA: Common Nano-Materials and Their Use in Real World Applications, Science Progress, Vol. 95, No. 1, pp. 1-22, 2012

[3] Lan W, Kexing F, Liang Y, Botao W: The Application of Ceramic Coatings in Petroleum Chemical and Building Industries, International Conference on Material and Environmental Engineering, 21.03-24.03.2014, Jiujiang, Jiangxi, China, Atlantis Press, ISBN 978-94-6252-004-2, pp. 146-149, 2014

[4] Orbán J: The Use of Nanotechnology in the Building Industry Part I. (in Hungarian), Magyar Építéstechnika, No. 1, pp. 40-43, 2012

[5] Orbán J: The Use of Nanotechnology in the Building Industry Part II. (in Hungarian), Magyar Építéstechnika, No. 2-3, pp. 54-57, 2012

[6] Abdelrahman M: Towards Sustainable Architecture with Nanotechnology, Al-azhar Engineering 11th International Conference, Cairo, paper 154, 2010

[7] Austrotherm Ltd.: Applications – Thermal Insulation Over Rafters: Manzard Grafif® (in Hungarian), Austrotherm Ltd., Győr, Hungary, p. 24

[8] Plundstein M: Dämmstoffarten. Detail Praxis – Dämmstoffe (Grundlagen, Materialen, Anwendungen), Institut für internationale Architektur-Dokumentation GmbH & Co. KG, München, Germany, pp. 17-57, 2007

[9] Fricke J, Heinemann U, Ebert HP: Vacuum Insulation Panels – From Research to Market, Vacuum, Vol. 82, No. 7, pp. 680-690, 2008

[10] Fricke J, Schwab H, Heinemann U: Vacuum Insulation Panels – Exciting Thermal Properties and Most Challenging Applications, International Journal of Thermophysics, Vol. 27, No. 4, pp. 1123-1139, 2006

[11] Simmler H, Brunner S: Vacuum Insulation Panels for Building Application: Basic Properties, Aging Mechanisms and Service Life, Energy and Buildings, Vol. 37, No. 11, pp. 1122-1131, 2005

[12] Schmidt M, Schwertfeger F: Applications for Silica Aerogel Products, Journal of Non-Crystalline Solids, Vol. 225, pp. 364-368, 1998
Baetens R, Jelle BP, Gustavsen A: Aerogel Insulation for Building Applications: A State-of-the-Art Review, Energy and Buildings, Vol. 43, No. 4, pp. 761-769, 2011

[13] Schultz JM, Jensen KI, Kristiansen FH: Super Insulating Aerogel Glazing, Solar Energy Materials and Solar Cells, Vol. 89, No. 2-3, pp. 275-285, 2005

[14] Hoffman D, Roy R, Komarneni S: Diphasic Ceramic Composites Via a Sol-Gel Method, Materials Letters, Vol. 2, No. 3, pp. 245–247, 1984

[15] Fullisol Ltd.: TSM Ceramics Thermal Insulation Coatings (in Hungarian), Fullisol Ltd., Budapest, Hungary, p. 10.

[16] Mart Ltd.: New Generation Thermal Insulating Nano-Coating with Micro-Sized Ceramic Balls, Mart Ltd., Dunakeszi, Hungary, p. 8.

[17] Thermo-Shield Inc.: Exterior Wall Coats, SPM Thermo-Shield Inc., Naples, Florida, United States

[18] MANTI Ltd.: Nanoceramic Thermal Protection (in Hungarian), MANTI Ltd., Budaörs, Hungary, p. 8.

[19] Paul G, Chopkar M, Manna I, Das PK: Techniques for Measuring the Thermal Conductivity of Nanofluids: A Review, Renewable and Sustainable Energy Reviews, Vol. 14, No. 7, pp. 1913-1924, 2010

[20] Nagy B: Laboratory Measurements of Construction Materials for Dynamic Heat and Moisture Transport Modelling (in Hungarian), In: S. Bodzás (ed.): Műszaki Tudomány az Észak-Kelet Magyarországi Régióban 2015, Debrecen (Hungary), Debreceni Akadémiai Bizottság Műszaki Szakbizottsága , ISBN:978-963-7064-32-6), pp. 446-452, 2015

[21] Bozsaky D: Laboratory Test with Liquid Nano-Ceramic Thermal Insulation Coating, In: M. Hajdú, M. J. Skibniewski (ed.): Creative Constructions Conference 2015, Cracow (Poland), ISBN 978-963-269-491-7, pp. 81-86, 2015

Acta
Technica
Jaurinensis

Vol. 9, No. 1, pp. 42-53, 2016
DOI: 10.14513/actatechjaur.v9.n1.394
Available online at acta.sze.hu

Measuring the Efficiency of Parallel Discrete Event Simulation in Heterogeneous Execution Environments

G. Lencse[1], I. Derka[1]

[1]Széchenyi István University, Department of Telecommunications
Egyetem tér 1., 9026 Győr, Hungary
Phone: +36 96 503 400
e-mail: steve@sze.hu

Abstract This paper deals with the experimental testing and verification of the earlier proposed load balancing and coupling factor criteria for the conservative parallel discrete event simulation in heterogeneous execution environments whether they can ensure a good speedup. The definition of the relative speedup is extended to the heterogeneous systems in a natural way. This definition is used to measure the efficiency of the parallel simulation executed by heterogeneous systems. A closed queueing network is used as the simulation model. It is demonstrated on the example of a heterogeneous system containing 134 CPU cores of 6 different types of computers that a good speedup can be achieved using the load balancing and coupling factor criteria. It is shown that the extension of the definition of the relative speedup to the heterogeneous systems made it easy to judge the speedup of parallel discrete event simulation in heterogeneous execution environments.

Keywords: parallel discrete event simulation, heterogeneous execution environment, conservative synchronisation method, load balancing criterion, coupling factor criterion

1. Introduction

Event-driven discrete event simulation (DES) is a powerful method for the performance analysis of information and communication technology (ICT) systems. The detailed modelling and simulation of these systems often requires a huge amount of computing power and memory. Parallelisation can be a natural solution. Kunz [1] points out that as the ongoing development in the hardware sector favours an increasing number of processing units over an increasing speed of a single unit thus the parallel simulation will remain an important and active field of research.

However, because of the algorithm of the event-driven DES, parallel discrete event simulation (PDES) it is not an easy task and the achievable speedup is often limited. When doing PDES, the model of the system is divided into partitions (called Logical Processes), and the partitions are assigned to processors that are executing them. To maintain causality, the virtual times of the partitions must be synchronised. There are different methods for synchronisation [1]. The conservative method ensures that causality is never violated. An event can be executed only if we are certain that no events with smaller timestamp exist (and also will not be generated) anywhere in the model. Unless the simulated system has a special property that the so called lookahead is large enough, the processors executing the partitions need to wait for each other in the majority of time, so the achievable speedup is poor.

A method for assessing available parallelism in a simulation model for conservative synchronization was proposed in [2]. The method requires only a small number of parameters that can be measured on a sequential simulation easily. The results of the aforementioned work were checked for homogeneous clusters up to 24 CPU cores in [3] and it was also examined how the different parameters of the model influence the achievable speedup. Criteria for a good speedup in a heterogeneous execution environment were proposed in [4]. The criteria were justified by several measurements in a test system. Further results on the examinations of different factors that influence the achievable speedup of parallel discrete event simulation in a heterogeneous execution environment were presented in [5]. Moreover, the definition of the relative speedup was extended to heterogeneous systems and this extension was used in the discussion of the results the experiments to evaluate the efficiency of parallel simulation executed by heterogeneous systems in the same conference paper. In our current paper, conference paper [5] is extended with experiments using significantly higher number of CPU cores (however the different kinds of experiments published in that paper are not included in this paper due to space limitations).

The remainder of this paper is organised as follows: first, a brief summary of the method for assessing the available parallelism is given. Second, our concept of heterogeneous execution environment, our criteria for a good speed up and our previous results are summarized. Third, the definition of the relative speedup is extended to the heterogeneous systems to be able to express the efficiency of parallel simulation executed by heterogeneous systems. Fourth, our heterogeneous test environment and simulation model are described. Fifth our measurements taken on the largest heterogeneous cluster are presented and discussed. Sixth, super-linear speedup is demonstrated on a large homogeneous cluster. Finally, our paper is concluded.

This topic was identified as being of importance in the parallel simulation of large systems using heterogeneous execution environments.

2. Method for Assessing Available Parallelism

The available parallelism can be assessed using some quantities that can be measured during a sequential simulation of the model in question. The following description is taken from [3].

The paper [2] uses the notations *ev* for events, *sec* for real world time in seconds and *simsec* for simulated time (model time) in seconds. The paper uses the following quantities for the assessing of available parallelism:

- P performance represents the number of events processed per second (*ev/sec*).
- E event density is the number of events that occur per simulated second (*ev/simsec*).
- L lookahead is measured in simulated seconds (*simsec*).
- τ latency (*sec*) is the latency of sending a message from one Logical Process (LP) to another.
- λ coupling factor can be calculated as the ratio of LE and τP:

$$\lambda = \frac{L \cdot E}{\tau \cdot P} \tag{1}$$

It was shown in [3] that if λ is in the order of several hundreds or higher then we may expect a good speedup. It may be nearly linear even for higher number of segments (N) if λ_N is also at least in the order of several hundreds, where:

$$\lambda_N = \frac{\lambda}{N} \tag{2}$$

3. Modelling and Simulation in Heterogeneous Execution Environments

This chapter is a summary of [4].

3.1. Our Concept of Heterogeneous Execution Environments

A logical topology of two levels was proposed: a *star shaped network* of *homogeneous clusters*. This model is simple enough and can describe a typical *heterogeneous execution environment*. What is logically described as a homogeneous cluster, it can be physically, for example, a cluster of PCs with identical configuration interconnected by a switch or it can be a chassis based computer built up by several main boards, etc. The main point is that a homogeneous cluster is built up by identical configuration elements especially concerning CPU type and speed as well as memory size and speed. The homogeneous clusters are interconnected logically in a star shaped topology. The physical connection can be a switch or the topology may be different but our model considers it to be a star for simplicity.

3.2. Criteria for Achieving a Good Speedup

Two criteria were set up. The ***load balancing criterion*** requires that all the CPUs (or CPU cores) should get a fair share from the execution of the simulation. A fair share is proportional to the computing power of the CPU *concerning the execution of the given simulation model*. (This is very important, because, for example, using different benchmark programs for the same set of computers one can get seriously different performance results.) Thus, for the fair division of a given simulation model among the CPUs, the CPUs should be benchmarked by the same type of simulation model that is to be executed by them (but smaller in size, of course). The *lookahead* or ***coupling factor criterion*** is the same as presented and tested in [3] up to 24 CPU cores.

3.3. Most Important Results

The load balancing criterion was justified by measuring the execution time of a model with different partitioning. The results of our experiments were quite close to the values computed according to the load balancing criterion. The coupling factor criterion was justified by different scenarios including a simulation executed by 64 CPU cores of 4 types resulting in a good speedup in [4].

4. Efficiency of Parallel Simulation Executed by Heterogeneous Systems

4.1. Relative Speedup of Program Execution by Heterogeneous Systems

First, the definition of the relative speedup of parallel execution of programs is extended for heterogeneous systems (in general, not only for simulation).

The conventional definition of the *speedup* (s_n) of parallel execution is the ratio of the speed of the parallel execution by *n* CPUs and the sequential execution by 1 CPU that is equal with the ratio of the execution time of the sequential execution (T_1) and that of the parallel execution (T_n):

$$s_n = \frac{T_1}{T_n} \tag{3}$$

The *relative speedup* (r_n) can be calculated as the ratio of the speedup and the number of the CPUs that produced the given speedup:

$$r_n = \frac{s_n}{n} \tag{4}$$

The relative speedup measures the *efficiency* of parallel execution. A relative speedup value of 1 means that the speedup is linear that is the computing power of *n* CPUs can be fully utilized.

When dealing with heterogeneous systems, not only the number of the CPUs but also their performance is to be taken into consideration. We were looking for a definition of the relative speedup of heterogeneous systems that can be used to measure the efficiency of program execution by the heterogeneous systems in the same way: its value of one should mean that the computing power of all the CPUs (from different types) can be fully utilized.

Let us denote the *number of the CPU types* in a heterogeneous system by *NT*, the *number* and the *performance* of CPUs available from *type i* by N_i and P_i, respectively. The *cumulative performance* of the heterogeneous system is:

$$P_c = \sum_{j=1}^{NT} P_i \cdot N_i \tag{5}$$

Let us denote the *execution time* of a given program by a single CPU of *type i* by T_i and let T_h denote the *execution time* of the given program by the *heterogeneous system*. The speedup of the heterogeneous system compared to the sequential execution by one CPU of type *i* is:

$$s_{h/i} = \frac{T_i}{T_h} \tag{6}$$

The relative speedup of the heterogeneous system against the sequential execution by one CPU of type *i* is now defined as:

$$r_{h/i} = \frac{T_i \cdot P_i}{T_h \cdot P_c} \tag{7}$$

We believe that this definition can be used in general for measuring the efficiency of program execution by heterogeneous systems. Thus, for simplicity, we used the word “CPU” in this section, but the expression “CPU core” could be used instead, as it is used everywhere else in this paper.

4.2. Measuring the Efficiency of Parallel Simulation Executed by Heterogeneous Systems

If the above definition of the relative speedup of program execution by heterogeneous systems is used for measuring the efficiency of parallel simulation executed by heterogeneous systems and the performance values of the CPU cores from different types are measured by benchmarking them with the same simulation model (expressing its value in events per seconds and the value of execution time in seconds) then the numerator of expression (7) gives the same values for all the values of *i* (that is its value is independent of the CPU core types): it is equal with the total number of events in the sequential simulation.

Note that the number of events in the parallel version of the simulation may be higher (e.g. due to communication and synchronization overhead) but only the events of the original sequential simulation are the essential part of the operation of the original model. Thus efficiency should consider the events of the original sequential simulation only.

Denoting the total *number of events* in the sequential simulation by N_E, definition (7) can be rewritten as:

$$r_h = \frac{N_E}{T_h \cdot P_c} \tag{8}$$

Thus, we have shown that the value of relative speedup calculated by (8) does not depend on which CPU core it was calculated against, it characterises the parallel simulation itself. Note that this is still true if one uses definition (7), thus one can use any of them selecting on the basis of which values are easier to measure directly in the given simulation.

5. Heterogeneous Test Environment

5.1. Available Hardware Base

The following servers, workstations and PCs were available for our experiments. Note that this hardware base is significantly larger than that of [5].

One Sun Server SunFire X4150 (referred as: BigSun)

- Two Quad Core Intel Xeon 2.83GHz CPUs
- 8GB DDR2 800MHz RAM
- Two near-line SAS 160GB HDD
- Two Intel 82571EB Gigabit Ethernet NIC
- Two Intel 80003ES2LAN Gigabit Ethernet NIC

Altogether it means a homogeneous cluster of 8 cores.

Thirteen LS20 Blades in an IBM BladeCenter (referred as: IBM)

- Two Dual Core Opteron 280 2.4GHz CPUs
- 4GB DDR2 667MHz RAM
- 73GB SCSI Ultra 320 HDD
- Broadcom NetXtreme BCM5704S Gigabit Ethernet NIC

Altogether it means a homogeneous cluster of 52 cores.

Ten Dell Precision 490 Workstations (referred as: Dell)

- Two Intel Xeon 5140 Dual Core 2.33GHz CPUs
- 4x1GB DDR2 533MHz RAM (quad channel)
- 80GB SATA HDD
- Four Intel 82571EB Gigabit Ethernet NICs
- Broadcom NetXtreme BCM5752 Gigabit Ethernet NIC

Altogether it means a homogeneous cluster of 40 cores.

Eleven AMD PCs (referred as: AMD)

- AMD Athlon 64 X2 Dual Core 4200+ 2.2GHz CPU
- 2GB DDR2 667 MHz RAM
- Two 320 GB SATA HDD
- nVidia CK804 Gigabit Ethernet NIC

Altogether it means a homogeneous cluster of 22 cores.

One Dell Precision 490 Workstation with faster CPUs (referred as: FastDell)

- Two Intel Xeon 5160 Dual Core 3GHz CPUs
- 4x1GB DDR2 533MHz RAM (quad channel)
- 80GB SATA HDD
- Four Intel 82571EB Gigabit Ethernet NICs
- Broadcom NetXtreme BCM5752 Gigabit Ethernet NIC

Altogether it means a homogeneous cluster of 4 cores.

Two Sun Servers SunFire X4200 (referred as: LittleSun)

- Two 2.2GHz Dual Core AMD Opteron 275 CPUs
- 4x1GB 400MHz ECC SDRAM
- 80GB HDD
- Four Intel 82546EB Gigabit Ethernet NICs

Altogether it means a homogeneous cluster of 8 cores.

Switches for interconnection:

- Two 3Com Baseline 2948 SFP Plus 48 ports Switches (3CBLSG48)
- Cisco Intelligent Gigabit Ethernet Switch Module, 4 ports (Part Number 32R1894) in the BladeCenter

5.2. Software Environment

Operating Systems

Debian Squezee (x86_64) Linux was used on all the computers.

Cluster Software

OpenMPI 1.6.2 (x86_64)

Discrete-Event Simulation Software

The widely used, open source OMNeT++ 4.2.2 discrete-event simulation environment [6] was chosen. It supports the conservative synchronization method (the Null Message Algorithm) since 2003 [7] We also expect that because of the modularity, extensibility and clean internal architecture of the parallel simulation subsystem, the OMNeT++ framework has the potential to become a preferred platform for PDES research.

5.3. The Simulation Model

Bagrodia and Takai proposed the Closed Queueing Network for testing the performance of conservative parallel discrete-event simulation [8]. OMNeT++ has a CQN implementation among its simulation sample programs. We have found this model perfect for our purposes and it was used in our current paper as well as in [3]-[5]. The below description of the model is taken from [3].

This model consists of *M* tandem queues where each tandem consists of a switch and *k* single-server queues with exponential service times (Fig. 1, left). The last queues are looped back to their switches. Each switch randomly chooses the first queue of one of the tandems as destination, using uniform distribution. The queues and switches are connected with links that have nonzero propagation delays. The OMNeT++ model for CQN wraps tandems into compound modules.

To run the model in parallel, we assign tandems to different LPs (Fig. 1, right). Lookahead is provided by delays on the marked links.

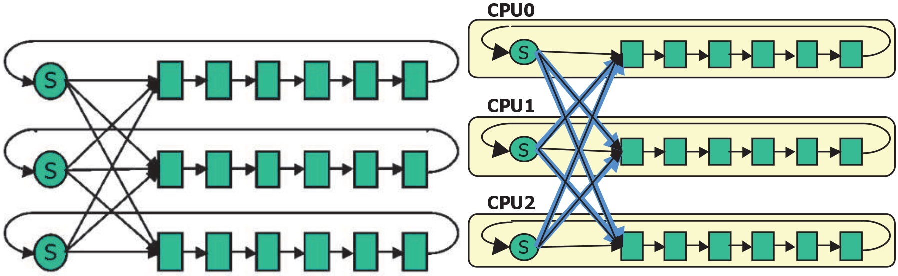


Figure 1. M=3 tandem queues with k=6 single server queues in each tandem queue – and its partitioning (taken from [3])

As for the parameters of the model, the preset values shipped with the model were used unless it is stated otherwise. Configuration B was chosen, the one that promised good speedup.

6. Experiments and Results of our Largest Heterogeneous Cluster

Fig. 2 shows the interconnection of the elements of our heterogeneous environment.

The automatic CPU frequency scaling [9] was switched off on all the computers by setting the CPU clock frequency to its maximum value. Unlike before, now Debian Squeeze (x86_64) operating system was installed on all the computers.

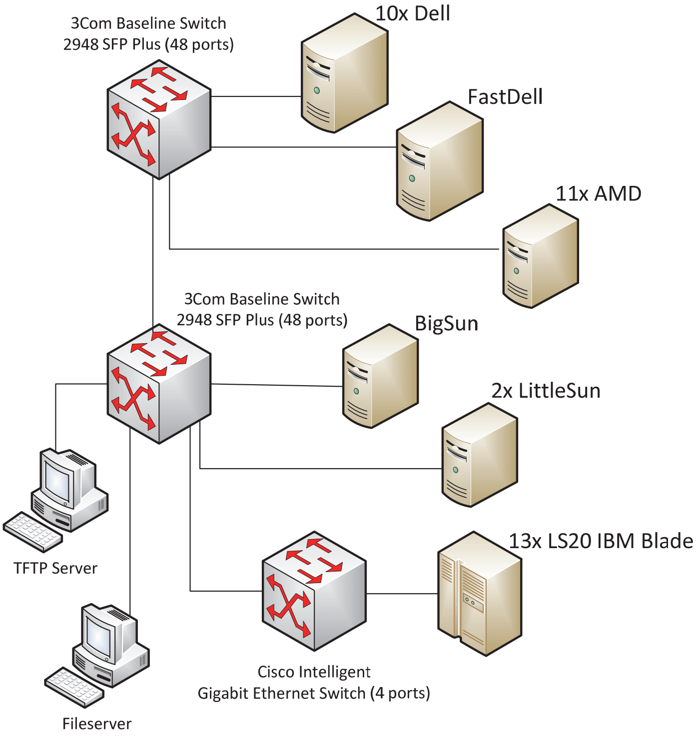


Figure 2. The elements of the heterogeneous execution environment with their interconnections

The number of the single queues in the tandem queues was also revised compared to papers [4] and [5] for the following reasons. In paper [3], M=24 tandem queues were used having each k=50 queues in them. Since then, the number of tandem queues were increased up to 480 (so that they can be distributed among the high number of cores) but the number the single queues were kept. This resulted in a significant change in the ratio of the two values. Now, the original ratio was restored by using the following values: M=480, k=1000. L=2000s was kept from the last test in [5]. As the expected value of the service time of the single queues was kept 10 seconds, the effect of the vacationing jobs (see [3]) was significantly reduced, because now the events could spend much more time in the tandem queues then in the long delay lines.

All the CPUs were benchmarked using a smaller network with the following parameters: M=24, k=1000, L=2000*sec* and the test were executed until 10^4*simsec*. All the tests were executed 11 times and average and standard deviation were calculated. The results are shown in Table 1. These performance values were used for the performance proportional partitioning of the simulation model containing 480 tandem queues among the 134 cores of our largest heterogeneous cluster. The partitioning is shown in Table 2.

Table 1. The Benchmarking Results (ev/sec) for the Largest Heterogeneous Cluster

Core Type	***BigSun***	***IBM***	***Dell***	***AMD***	***FastDell***	***LittleSun***
Average	264 687	176 118	205 116	173 764	240 491	173 540
Std. Dev.	1 762	1 221	1 631	1 686	1 472	1 768

Table 2. The Division of the 480 Tandem Queues Among the Cores of the Largest Heterogeneous Cluster

Core type	***P_i (ev/sec)***	***N_i***	***n_i***	***no. of cores***	***tandems /core***	***cumulated tandems***
BigSun	264 687	8	4.95	8	5	40
IBM	176 118	52	3.30	38	3	114
				14	4	56
Dell	205 116	40	3.84	40	4	160
AMD	173 764	22	3.25	22	3	66
FastDell	240 491	4	4.50	4	5	20
LittleSun	173 540	8	3.25	8	3	24
No. of all the cores:		134	Number of all the tandems:			480

In order to keep the total sum of number of the tandem queues exactly 480, we could not always follow the rules of the mathematical rounding: 4 tandem queues were assigned to each one of 14 CPU cores from among the 52 cores of the IBM blades whereas only 3 tandem queues were assigned to each one of the other 38 CPU cores from the same type (which one matches with the result of the correct rounding of 3.3).

Using the 134 cores, 10^5*simsec* long simulations of the CQN model were executed eleven times and average and standard deviation of the execution time were calculated. The average execution time was 278.19*sec* with the standard deviation of 2.63*sec* (which is less than 1% of the average).

For the calculation of the relative speedup we need either N_E the total number of events in the sequential execution of the simulation or the execution time of the sequential simulation using one of the CPU cores. We executed the sequential simulation of the same model and number of events was: N_E=6.1675*10^9*ev*. Using the cumulative sum of the performance of 134 cores, P_c=25.653*10^6*ev/sec*, the relative speedup of the heterogeneous cluster is:

$$r_h = \frac{6.1675 \cdot 10^9 ev}{278.19 sec \cdot 25.653 \cdot 10^6 ev/sec} \approx 0.8642 \qquad (9)$$

This is an excellent result produced with 134 CPU cores of 6 different types of computers. For comparison, the relative speedup of the heterogeneous cluster was 0.88 using only 86 CPU cores from 4 types in [5].

7. Testing a Large Homogeneous Cluster

Finally, we tested a large homogeneous cluster. Even though we had 13 blades, we used only 12 of them (having 48 cores) to be able to evenly divide the 480 tandem queues. The parameters of the model were kept from the previous experiment. The average execution time of the 11 simulation runs was 730.68*sec* with 3.22*sec* standard deviation (which is 0.44% of the average) whereas the execution time of the sequential simulation performed by a single IBM Blade core was 52077*sec* (!). This means a 71.27 speedup produced by a system having only 48 core with is about 1.5 relative speedup. This phenomenon of the super linear speedup can be easily explained by the effect of caching: the working set of the simulation program can better fit into the cache of the 48 CPU cores than into that of a single CPU core. Similar phenomenon was reported in [10], see page 95.

8. Conclusion

An extension of the relative speedup for heterogeneous systems was introduced. It proved to be an appropriate tool for the evaluation of the speedup of parallel discrete event simulation in heterogeneous execution environments. We could achieve 0.86 relative speedup using a heterogeneous cluster built up by 134 CPU cores of 6 different types of computers. We also demonstrated a super-linear speedup (relative speedup of 1.5) on a 48 core homogeneous cluster.

References

[1] Kunz G: Parallel discrete event simulation, in K. Wehrle, M. Günes and J. Gross (Eds.) Modeling and Tools for Network Simulation, Berlin, Springer-Verlag, pp. 121-131, 2000

[2] Varga A, Sekercioglu YA, Egan GK: A practical efficiency criterion for the null message algorithm, in Proc. 15th European Simulation Symposium, Delft, Netherlands, pp. 81-92, 2003

[3] Lencse G, Varga A: Performance prediction of conservative parallel discrete event simulation, in Proc. 2010 Industrial Simulation Conference, Budapest, Hungary, pp. 214-219, 2010

[4] Lencse G, Derka I, Muka L: Towards the efficient simulation of telecommunication systems in heterogeneous execution environments, in Proc. 36th International Conference on Telecommunications and Signal Processing, Rome, Italy, pp. 304-310, 2013

[5] Lencse G, Derka I: Testing the speedup of parallel discrete event simulation in heterogeneous execution environments, Proc. 11th Annual Industrial Simulation Conference, Ghent, Belgium, pp. 101-107, 2013
[6] Varga A, Hornig R: An overview of the OMNeT++ simulation environment, Proc. 1st International Conference on Simulation Tools and Techniques for Communications, Networks and Systems & Workshops, Marseille, France, 2008, http://dl.acm.org/citation.cfm?id=1416290
[7] Sekercioglu YA, Varga A, Egan GK: Parallel simulation made easy with OMNeT++, Proc. 15th European Simulation Symposium, Delft, The Netherlands, pp. 493-499, 2003
[8] Bagrodia RL, Takai M: Performance evaluation of conservative algorithms in parallel simulation languages, IEEE Transactions on Parallel and Distributed Systems, Vol. 11, No. 4, pp. 395-411, 2000
[9] Archlinux, CPU frequency scaling, https://wiki.archlinux.org/index.php/CPU_frequency_scaling
[10] Benzi J, Damodaran M: Parallel three dimensional direct simulation Monte Carlo for simulating micro flows, in Parallel Computational Fluid Dynamics 2007, Springer, 91-98, 2009
DOI: 10.1007/978-3-540-92744-0_11

Acta
Technica
Jaurinensis

Vol. 9, No. 1, pp. 54-64, 2016
DOI: 10.14513/actatechjaur.v9.n1.396
Available online at acta.sze.hu

Spectral Reconstruction with Genetic Optimization in Case of Different Sample Sets

Zs. Sávoli[1], A Horváth[2], B. Kránicz[1]

[1]Interdisciplinary Doctoral School of Engineering, Széchenyi István University, Egyetem tér 1., 9026, Győr, Hungary, e-mail: savoli@mail.barossgyor.hu

[2]Department of Physics and Chemistry, Széchenyi István University, Egyetem tér 1., 9026, Győr, Hungary, e-mail: horvatha@sze.hu

Abstract: Finding the spectral features of color sample sets is a main issue of colorimetry. It is widespread to apply Principal Component Analysis in these researches. Several studies were written about the reconstruction of samples with known spectra using principal components, and some works dealt with approximate reconstruction from tristimulus values. Our study examines how spectral reconstruction done with genetic optimization works in case of different sample sets. We have taken only the tristimulus values of the color samples given.

Keywords: spectral reconstruction, tristimulus values, principal component analysis, genetic optimization, different samle sets

1. Introduction

It is a basic thesis of colorimetry that any color stimulus can unequivocally be given by three numbers. These numbers can be the CIE X, Y, Z tristimulus values or the coordinates of any other suitable color space ($Y, x, y; L^*, a^*, b^*; L^*, u^*, v^*$ etc.). However, the description of selfluminous objects (light sources) and surfaces (secondary light sources) is often given with spectral features, in other words with spectra. This type of description provides much more information about the observed object, it needs, therefore, more than three parameters. The researchers in colorimetry started examining how they can determine or give approximately the reflection spectra of surfaces with only a few numbers. Principal Component Analysis (PCA), which is based on elements of mathematical statistics and linear algebra, has appeared to be an especially strong and interesting tool.

Several studies deal with the usage of Principal Component Analysis in colorimetry, therefore its mathematical presentation is not the subject of the present article. Publications [1]-[7] provide a general view how it works. In order to apply this method effectively, it is necessary to have a set with a large number of known spectra. Principal

Component Analysis produces the eigenvectors belonging to the sample set. The linear combination of these eigenvectors helps to reconstruct the spectrum. The above mentioned studies present reconstruction of samples that have known and measured spectra. The question is whether it is possible to say anything about the spectrum of color samples that have unknown reflectance functions if we only know their tristimulus values X,Y,Z. In some of the studies focusing on this problem [8], [9], the spectral reconstruction is done with an algorithm using pseudoinvers matrix operations.

However, Principal Component Analysis or weighed Principal Component Analysis has been used in other studies. Corresponding to the tristimulus values, the three eigenvectors, whose eigenvalues are the greatest ones, are enough to get the same X,Y,Z values as the result of reconstruction [10], [11]. More accurate results can be reached with more components, but they lead to undetermined equation systems, as the tristimulus values can be given in many ways from more than three components. Because of this critical statement, three vectors have been considered to be satisfactory in most cases.

Publication [12] describes our method that deals with this problem. We used the first five eigenvectors instead of the first three vectors and the features of the reflection functions of real color samples for generic optimization. It was considered that the reflection functions of the real samples are smooth, without strong oscillation and they are non-negative. This method provides a much more accurate reconstruction than the earlier ones. (Figure 1) The set of 2 832 textile samples with known reflectance functions were used. Furtheremore, 148 flower samples, 565 paint samples and 8 533 human skin samples were applied. Figure 2 shows these sample sets. The formalism of optimization is to be shown in brief. More details can be read in publication [12].

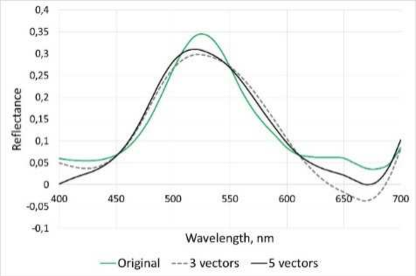
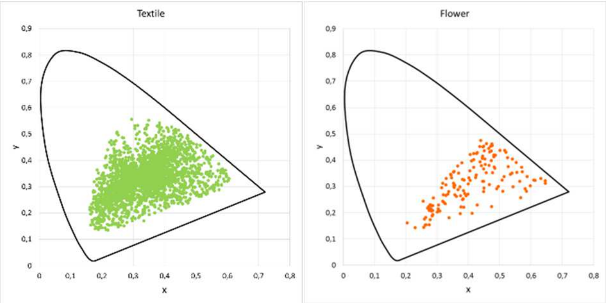


Figure 1. Reconstruction with three and with five eigenvectors



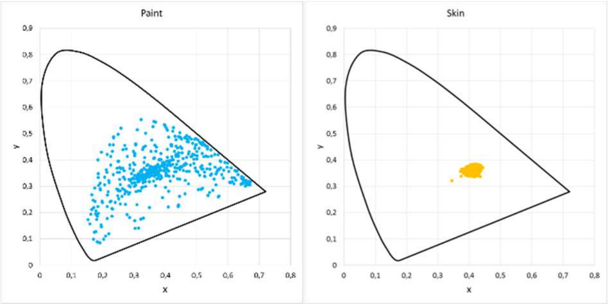


Figure 2. The textile, flower, paints, skin samples sets in CIE xy color space

2. The formalism of optimization

During calculation, we work with spectra (reflectance functions) whose resolution is given, therefore, we use finite-dimensional vectors instead of continuous functions. Let N denote the dimension number of these vectors. For example, if we think of a spectrum with a range of 400 nm - 700 nm and with an equidistant-wavelength step of at 10 nm, $N = 31$.

Let us denote the eigenvalues of the PCA method arranged in decreasing order by $\tau_1 \geq \tau_2 \cdots \tau_N \geq 0$, the eigenvectors relating to the eigenvalues by $v_1, v_2, \cdots, v_N$, and the mean vector by m. The linear combination of M eigenvectors and the principal components $c_1, c_2, \cdots, c_M$ provides the following spectrum.

$$f(c_1, c_2, \cdots, c_M) = \sum_{i=1}^{M} c_i \cdot v_i + m, \tag{1}$$

$f(c_1, c_2, \cdots, c_M)$ is also an N-dimensional vector.

Having M fixed, variables $c_1, c_2, \cdots, c_M$ determine the spectrum of the reconstructed f according to equation (1). As a next step, we create a function that measures the difference between this type of spectrum and the ideal spectrum. The difference is small for smooth and non-negative metamers, and it is greater and greater, if the tristimulus values deviate from the stipulated ones or if the function oscillates strongly or it takes up negative values. It is easy to calculate the tristimulus values of the spectrum by the application of color-matching functions.

$$X_0 = \sum_{i=1}^{N} f_i \cdot S_i \cdot \bar{x}_i, Y_0 = \sum_{i=1}^{N} f_i \cdot S_i \cdot \bar{y}_i, Z_0 = \sum_{i=1}^{N} f_i \cdot S_i \cdot \bar{z}_i \tag{2}$$

In Eq. 2, $\bar{x}_i, \bar{y}_i, \bar{z}_i$ denote the discrete versions of the CIE color-matching functions which have the same resolution as that of the spectra, S_i .is the discrete spectral power distribution of the illuminant. Obviously, the values of X_0, Y_0, Z_0 depend on the coefficients c_i, but we do not emphasize this dependence for the sake of briefness.

We can calculate the squared sum of the differences to show how much the values X_0, Y_0, Z_0 deviate from the predefined values X, Y, Z.

$$d_0(c_1, c_2, \cdots, c_M) = (X - X_0)^2 + (Y - Y_0)^2 + (Z - Z_0)^2 \tag{3}$$

This d_0 value is non-negative and it is equal to 0 when a metamer complies with the definition. If $M = 3$, the equation system of the metamer has a single solution. A lot of earlier studies which used PCA ended with giving this solution. If $M > 3$, it has an infinite number of solutions, and we can choose the most realistic one by the use of constraints on negativity and strong oscillation.

We can describe the negativity of the function by the integral of the negative and the positive part, or with the ratio of their sums in the discrete case. Denote:

$$f_i^+ = \max(f_i, 0) \tag{4}$$

$$f_i^- = \min(f_i, 0) \tag{5}$$

$$F^+ = \sum_{i=1}^{N} f_i^+ \tag{6}$$

$$F^- = -\sum_{i=1}^{N} f_i^- \tag{7}$$

Definitions (4) and (5) resemble the terms of lower and upper covering functions used in analysis.

The penalty term on negativity is:

$$P_n = \frac{F^-}{(F^+ + F^-)} \cdot W_n \tag{8}$$

W_n is the weight factor which is used to set the relative weight of this term within the optimization function. It is obvious that $P_n = 0$ if the function has only non-negative values and $W_n = 1$, and the more negative parts, f contains the greater positive values P_n has got. It is 1 in extreme cases.

We define the oscillation of the function by the squared sum of the deviation between the neighboring terms:

$$V = \sum_{i=1}^{N-1} (f_{i+1} - f_i)^2 \tag{9}$$

The penalty term on oscillation is

$$P_v = \sqrt{\frac{V}{(N-1)}} \cdot W_v \tag{10}$$

The cost function whose minimum is assumed to determine the metamer with the best qualitative features is the following.

$$d(c_1, c_2, \cdots, c_M) = d_0 + P_n + P_v \tag{11}$$

We can get d_0 from (3), P_n from (8) and P_v from (10). The weights W_n and W_v show the importance of one or the other penalty terms. According to the pre-calculation we have already made, the useful values are $W_n = 100, W_1 = 1$. A little change in them will not influence the final result.

All in all, d is a non-linear function with M variables, whose minimum corresponds to the best function for us, in other words, vector $(c_1, c_2, \cdots, c_M)$ which gives the location of the extreme values, contains the optimal weight of the eigenvectors used in the reconstruction.

In order to find the minimum point of function d , which has been given in equation (11) above, we use our own genetic optimization program. The genetic algorithm was chosen because d has a lot of local minima (mainly because of the oscillation term) and the gradient-based methods generally cannot find the global minimum in these cases.

Our genetic algorithm uses the standard genetic operators, e.g. mutation and crossing, and in order to accelerate the search for local maxima, it uses hill-climbing steps. We had already applied this code to solve more industrial optimization problems [13].

3. Values describing the accuracy of the reconstruction

We can describe the accuracy of the reconstruction by several measuring numbers in a quantitative way. We can measure the color difference between the examined sample and the reconstructed sample, the sepctral deviation and accuracy between the original and the reconstructed reflection funtion.

The color difference can be given as it follows. As the first step, the tristimulus values X, Y, Z of the samples have to be transformed into values L^*, a^*, b^* (12), where X_n, Y_n, Z_n are the tristimulus values of the reference white tristimulus values under a given illuminant.

$$L^* = 116 \cdot f\left(\frac{Y}{Y_n}\right) - 16, a^* = 500 \cdot \left(f\left(\frac{X}{X_n}\right) - f\left(\frac{Y}{Y_n}\right)\right), b^* = 200 \cdot \left(f\left(\frac{Y}{Y_n}\right) - f\left(\frac{Z}{Z_n}\right)\right)$$

$$f(t) = \begin{cases} t^{\frac{1}{3}}, t > \left(\frac{6}{29}\right)^3 \\ \frac{1}{3} \cdot \left(\frac{29}{6}\right)^2 \cdot t + \frac{4}{29}, t \leq \left(\frac{6}{29}\right)^3 \end{cases} \tag{12}$$

In case of all the used samples, the condition $t > \left(\frac{6}{29}\right)^3$ is met. The X_n, Y_n, Z_n values of the reference white tristimulus under the illuminant with $S(\lambda)$ spectral power distribution are given by equation (13).

$$X_n = k_\varphi \cdot \int_{400}^{700} S(\lambda) \cdot \bar{x}_\omega(\lambda) d\lambda, Y_n = 100, Z_n = k_\varphi \cdot \int_{400}^{700} S(\lambda) \cdot \bar{z}_\omega(\lambda) d\lambda \tag{13}$$

$$k_\varphi = \frac{100}{\int_{400}^{700} S(\lambda) \cdot \bar{y}_\omega(\lambda) d\lambda} \tag{14}$$

The normalisation coefficient is denoted by k_φ and $\bar{x}_\omega(\lambda), \bar{y}_\omega(\lambda), \bar{z}_\omega(\lambda)$ are the CIE color matching functions. The values of the comparable color samples are L_1^*, a_1^*, b_1^*, and L_2^*, a_2^*, b_2^*. Knowing these, the CIE Lab ΔE_{ab}^* color difference can be determined (15).

$$\Delta E_{ab}^* = \sqrt{\left(L_2^* - L_1^*\right)^2 + \left(a_2^* - a_1^*\right)^2 + \left(b_2^* - b_1^*\right)^2} \tag{15}$$

If $\Delta E_{ab}^* = 0$, the test samples are the same. If $\Delta E_{ab}^* = 1$, it gives the just perceptible difference under a given illuminant.

Two different values are given for the spectral accuracy by the publications. One of them is a numerical value, GFC (goodness of fit coefficient) (16).

$$GFC = \frac{\sum_{i=1}^{31} \varphi(\lambda_i) \cdot \varphi_r(\lambda_i)}{\sqrt{\sum_{i=1}^{31} \varphi(\lambda_i)^2} \cdot \sqrt{\sum_{i=1}^{31} \varphi_r(\lambda_i)^2}} \tag{16}$$

$\varphi(\lambda_i)$ denotes the sample reflection function at wavelength λ_i. $\varphi_r(\lambda_i)$ denotes the reconstructed spectrum at λ_i. If $GFC = 1$, the reconstructed function is perfectly identical with the original one. Therefore the closer GFC gets to 1, the more accurate the spectral reconstruction is. The formula in (16) corresponds to the cosine value of the angle of two abstract vectors.

The other accuracy specifying value is RMS (root mean square), which gives the reconstruction error according to the differences between the original and the reconstructed spectra (17).

$$RMS = \sqrt{\frac{1}{31} \cdot \sum_{i=1}^{31} \left(\varphi(\lambda_{\mathrm{i}}) - \varphi_r(\lambda_{\mathrm{i}}) \right)^2} \tag{17}$$

The notations in equation (17) are the same as the notations in equation (16). The smaller the value is, the smaller the spectral error is and the greater the reconstruction accuracy is.

4. Reconstruction in case of textile samples

During calculation, we reconstructed some randomly-chosen textile samples with the above mentioned process and wanted to find the ideal coefficients c_i. During optimisation, we presumed the equienergetic $CIE\ E$ illuminant. Figure 3 shows the reconstruction of randomly chosen samples with five eigenvectors.

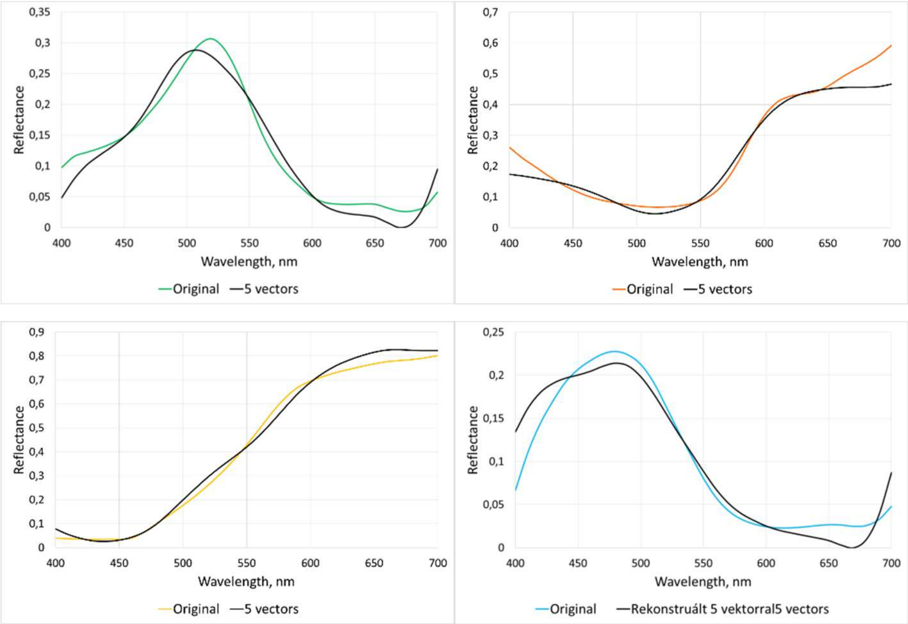


Figure 3. Reconstruction of textile samples. The black line denotes the reconstructed spectrum, the coloured line denotes the original spectrum

The following table (Table 1) shows the mean values of $\Delta E^*_{ab}, RMS, GFC$.

Table 1. The mean values $\Delta E^*_{ab}, RMS, GFC$ of the reconstruction done by genetic optimisation.

RMS	GFC	ΔE^*_{ab}
0.0300	0.9926	0.0100

5. Reconstruction with different sample sets

Let us examine what the coefficients and the spectra are if we use other samples than textiles. We suppose that different spectra belong to the same X, Y, Z tristimulus values in cases of other materials. The whole analysis is done for three other sample sets, namely, for 148 flowers, 565 paint samples and 8533 human skin samples, and the principal components c_i are determined with optimisation again.

However, the eigenvectors of PCA and the coefficients c_i are different in case of different sample sets, it is interesting to know, how big the difference is between the reconstructed spectra belonging to the same tristimulus values. Therefore, the reconstruction is done for randomly-chosen tristimulus values in case of different training sample sets. The reconstruction is done for the above mentioned sample sets, the textile samples, the flowers, the paint samples and the skin samples, with their own eigenvectors using our genetic oprimisation algorythm and presuming the illuminant $CIE\ E$. Figure 4 shows the reconstructed spectra of the different training sample sets belonging to the same tristimulus values X, Y, Z.

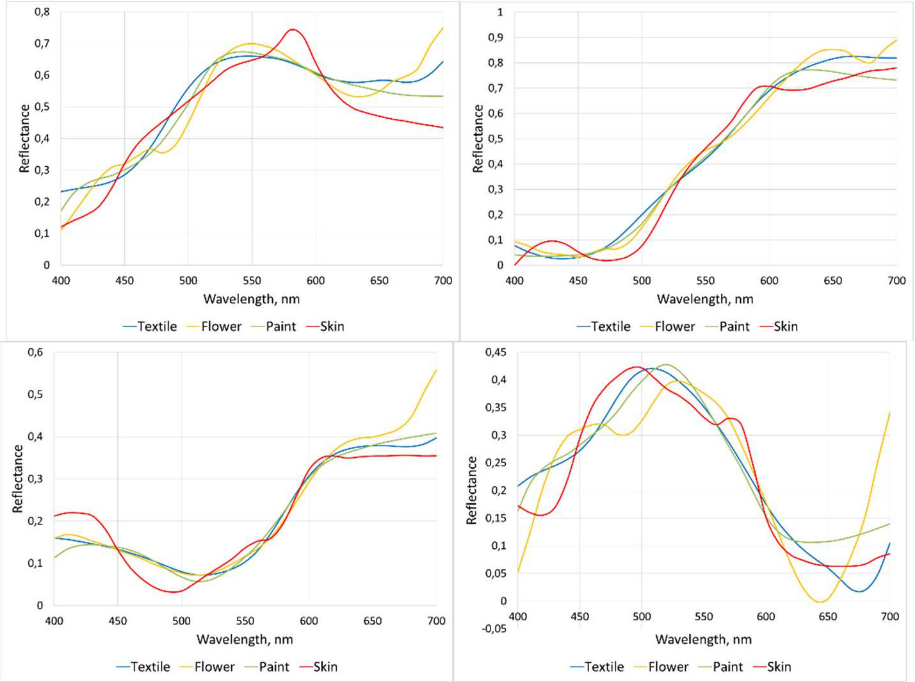


Figure 4. The reconstructed spectra belonging to given tristimulus values in case of different training sample sets

Figure 4 depicts clearly that different spectra belong to the same tristimulus values depending on the training sample sets. The reconstructed spectra are less oscillating for textile and paint samples than for flowers or human skin samples, which corresponds to the features of the samples. The reconstructed functions of the textile and paint samples resemble each other the most, however, they are not identical at all. Table 2 denotes the values of RMS, GFC which describe the reconstruction numerically, compared to the reconstructed spectrum. With other words, the basis of comparison is always the textile spectrum. The values ΔE_{ab}^{*} are not calculated as only the shapes of the reconstructed reflectance functions are compared.

Table 2. The mean values of RMS, GFC corresponding to the same tristimulus values fro different training sets

textile-flower		***textile-paint***		***textile-skin***	
RMS	GFC	RMS	GFC	RMS	GFC
0.0522	0.9821	0.0325	0.9963	0.0556	0.9847

The content of Table 2 depicts that different spectra belongs to the same tristimulus values when we use textile, flower, paint or skin samples. The differences are smaller in case of textile and paint samples than of other sample set pairs. This statement is proved for the first five eigenvectors and the mean vectors of each sample set. (Figure 5)

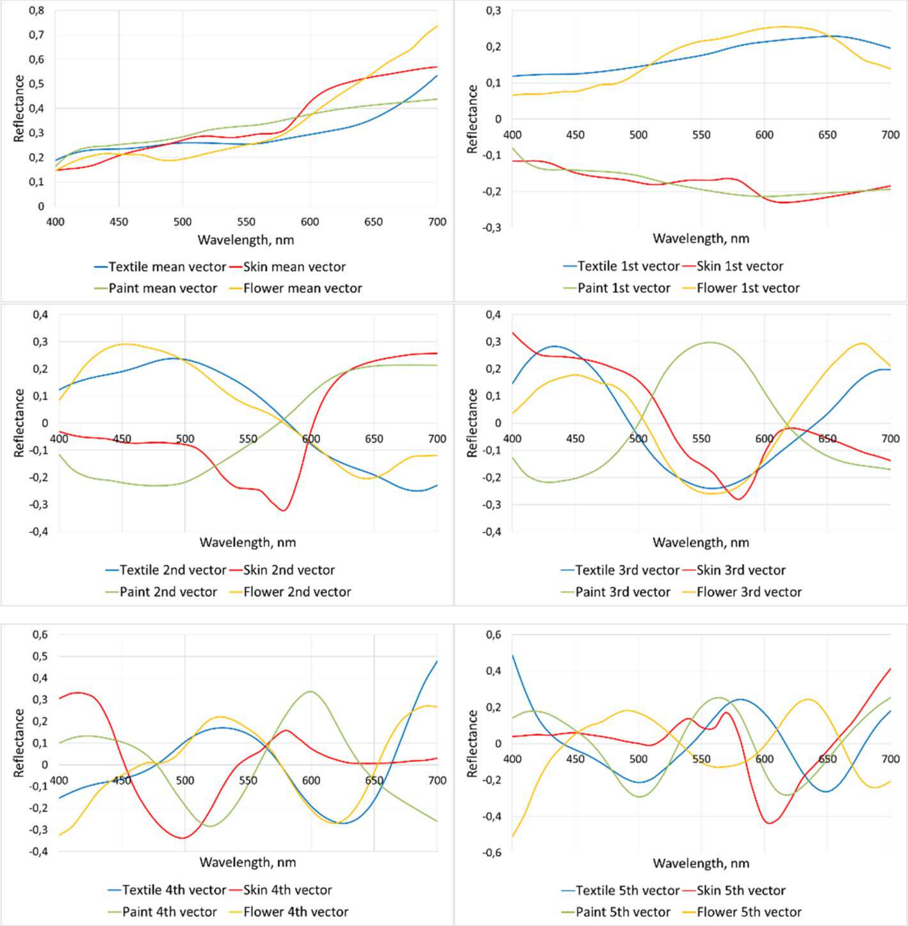


Figure 5. The mean vector and the first five eigenvectors for different sample sets

Figure 5 depicts that the eigenvectors of the textile and the paint samples differ from each other the least, on the other hand, the eigenvectors of the other two sample sets show greater differences. Therefore, different spectra belong to the same tristimulus values X, Y, Z for different sample sets.

Conclusion

Our study shows a genetic optimisation process that helps us to reconstrutct the reflectance funtion of colorful samples on the basis of their tristimulus values. During reconstruction we use the first five eigenvectors provided by the PCA together with restrictive terms for the shape of the reflectance functions of real samples. It gives the opportunity to choose those metamers from their infinite set in such a way that the reflectance function corresponds with the reflectance functions of the real samples.

All in all, our method makes it easy to reconstruct a spectrum for the tristimulus values of any colour sample with unknown spectrum. Even, we can tell what the spectrum looks like when the unknown sample is textile, flower, paint or skin. The method can be refined with the use of other sample sets.

References

[1] Tzeng DY, Berns RS: A Review of Principal Component Analysis and Its Application to Color Technology. Color Research And Application, Volume 30, Number 2, pp.84-98, 2005

[2] Maloney LT: Evaluation of linear models of surface spectral reflectance with small numbers of parameters. Journal of Optical Society of America, 3, pp. 1673–1683, 1986

[3] Jaaskelainen T, Parkkinen J, Toyooka S: Vector-subspace model for color representation. Journal of Optical Society of America, 7, pp. 725–730, 1990

[4] Vrhel MJ, Gershon R, Iwan LS: Measurement and analysis of object reflectance spectr. Color Research and Application, 19, pp.4–9, 1994

[5] García-Beltrán A, Nieves JL, Hernández-Andrés J, Romero J: Linear bases for spectral reflectance functions of acrylic paints. Color Research and Application, 23, pp.39–45, 1998

[6] Tajima J: A huge spectral characteristics database and its application to color imaging device design. Proceedings of the 6th IS&T/SID Color Imaging Conference, IS&T, Springfield, VA; pp. 86–89, 1998

[7] Hardeberg JY, Schmitt F, Brettel H: Multispectral color image capture using a liquid crystal tunable filter. Opt Engineering, 41, pp. 2533–2548, 2002

[8] Fairman HS, Brill MH: The principal components of reflectances. Color Research and Application, Vol. 29, No. 4, pp.104-110, 2004

[9] Harifi T, Amirshahi SH, Agahian F: Recovery of reflectance spectra from colorimetric data using principal component analysis embedded regression techniq. Opt. Rev., 15, pp. 302-308, 2008

[10] Farnaz A, Seyed AA, Seyed HA: Reconstruction of reflectance spectra using weighted principal component analysis. Color Research and Application, Vol. 33, No. 5, pp.360-371, 2008

[11] Simone B: Reflectance spectra recovery from tristimulus valuesby adaptiveestimation with metametric shape correction. Journal of Optical Society of America, Vol. 27, No. 8, pp.1868-1877, 2010
[12] Sávoli Zs, Horváth A, Kránicz B: Színminta-halmazhoz illeszkedő, jó kvalitatív tulajdonságú metamer gyors keresése. Képfeldolgozók és Alakfelismerők Társaságának 10. országos konferenciája, Kecskemét, pp 722-737, 2015
[13] Horváth A, Horváth Z: Optimal shape design of Diesel intake ports with evolutionary algorithm. M Feistauer (szerk.) Numerical Mathematics and Advanced Applications, ENUMATH 2003

Acta
Technica
Jaurinensis

Vol. 9, No. 1, pp. 65–82, 2016
DOI: 10.14513/actatechjaur.v9.n1.397
Available online at acta.sze.hu

Preprocessing Endoscopic Images of Colorectal Polyps

A. Horváth[1], Sz. Spindler[1], M. Szalai[2], I. Rácz[2]

[1]Széchenyi István University, Department of Physics and Chemistry
Győr, Hungary
[2]Petz Aladar County and Teaching Hospital, Győr, Hungary
E-mail: horvatha@sze.hu

Abstract: Classification of polyps in the human colorectum is a hot topic of gastroenterology. The current advancements in devices of endoscopy made it possible to product high resolution images of polyps with different light filters, e.g. with narrow band imaging (NBI) system. There exists several human classification methods that helps the doctor to decide whether a specific polyp is risky (i.e. it can be turn into cancer) or not. To overcome the limits of human classification skill, a digital image processing method was developed by the authors of this paper.

This paper summarizes the preprocessing method of this software, which is not a trivial task, because during the inspection of a patient, the diagnostician has limited control of light and geometry. Therefore the surface of the polyp is not uniformly illuminated, and prominent flares could appear on the images quite often. These circumstances makes it hard to detect the surface patterns precisely.

In this paper a lot of variants for this preprocessing task is presented. The best of them appears to reproduce the geometrical parameters of the patterns in 1–5% relative error which is enough for the classification method. The proposed method can be applied to another problems, where the precise measurement of patterns in highly variable illuminating conditions is needed.

Keywords: *image processing, thresholding, illumination*

1. Motivation

Colorectal cancer is one of the most lethal health problem of our age [1]. Many studies pointed out that this disease is responsible for almost the highest cancer related mortality rate in Europe. This is one of the reasons why clinical experts are working together with other field's specialists to find new diagnostic methods for this type of cancers, preferably in early stage.

Polyps are distortions of the human colorectum that can be detected as 1–20 mm diameter outgrowths the inner surface of the colorectum. Some type of polyps develop into cancer with high probability, meanwhile others remain polyps. In the former case endoscopic poly removal is recommended, in the latter case it is not necessary, if the size is not too big to form an obstacle in the colorectum.

The most accurate method of polyp classification is histology: colonoscopes are usually able to cut off small samples of suspicious part of the tissues (biopsy) and these samples can be examined later. This method has significant drawbacks: histology is time-consuming and requires specialist. Moreover, the biopsy itself generates approximately the same injury as the complete removal of the suspicious polyp. To circumvent these difficulties a lot of methods was developed and are under development to classify polyps based on endoscopic images only. In an ideal case, a good method can help to decide whether the polyp is dangerous (potentially can evolve into cancer) or not. Hence the patient have to bear the negative consequences of polyp ablation the procedure should be carried out only if it is really necessary.

Significant improvement in colonoscope hardware during the last decade made it possible to produce high resolution images and to choose some special light sources or filters. For example, the so called "Narrow Band Imaging" (NBI) filter light source into the 400 to 450 nm wavelength domain. In this domain the human blood absorbs the light with a very high efficiency, therefore the contrast among vessels and other tissues become more significant comparing to the white light illumination.(See Fig. 1 and [3].)

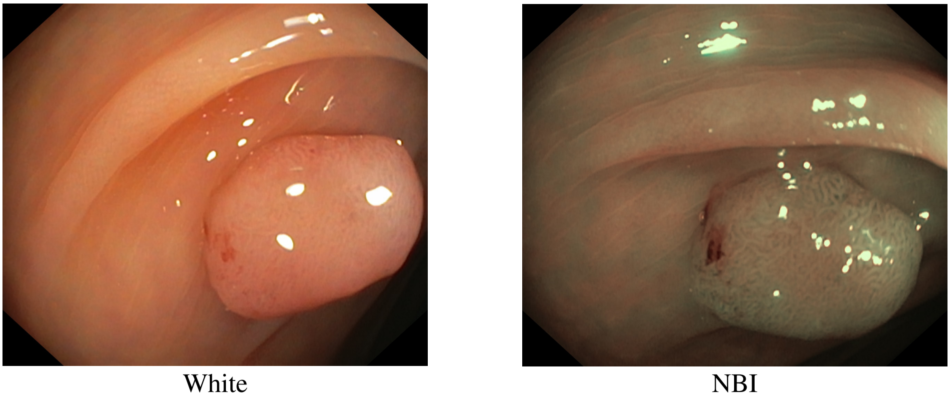

White NBI

Figure 1. Same polyp with white and NBI illumination

During the recent years several classification methods were developed for these high resolution and/or NBI images [4]. All of them agreed on using the surface patterns as a key feature but take other parameters into account too. It takes a lot of time and practice to detect and classify surface structures like polyps and different kind of lesions. As many studies pointed out the learning curve of classification is a very time consuming activity however there are specific trainings exist to facilitate the process. (See [6] and [7].)

The goal of this research is to develop a software that can decide whether a polyp can potentially evolve into cancer or not. In medical terminology doctors refer to the “harmless type” as hyperplastic (HP) and the others commonly called as non-hyperplastic (non-HP). The latter group can be divided into several sub classes, but it is out of our scope at this phase of the research.

2. Overview of the image processing method

To overcome the limits of human classification ability, a few research teams have been working on a method based on digital image processing.Széchenyi István University and Petz Aladár County Teaching Hospital also has a joint research in this field. The medical doctors in this project collected approximately 200 cases, with both high resolution NBI images and histological results. They classified these cases with standard Kudo and NICE methods. Authors of this paper developed a digital image processing software which can produce similar statistical accuracy that that of human classification methods. The main steps are:

1. Marking the surface of polyp in the picture. This is the “region of interest” (ROI) of the further investigations.
2. Preprocessing and noise filtering.
3. Identifying the small dark spots and measuring the geometrical parameters.
4. Classify the spots based on geometrical parameters with Quadratic Discriminant Analysis. (QDA)
5. Classify the polyp based on the statistics of spots and some global properties.

The whole method is too complex to publish it in one single paper, therefore only the 2nd and 3rd steps (preprocessing and identifying) will be described here in detailed. also several possible alternatives are going to be examined and the partly new type of adaptive thresholding method is presented.

The other steps are summarized here briefly.

2.1. ROI selection

Selection of the Region of Interest (ROI) is not trivial on endoscopic images. As it is clear from Fig. 1, it is not a simple task to a computer to find a polyp itself. Additional difficulty is that parts of the image can be distorted by humours, and prominent flare effects can arise. All of these areas are bad for processing and should not be included in ROI.

However the flares can be locked out automatically, the other problems would require too sophisticated algorithm. At this stage of the research human interaction is inevitable and for this reason an expert on this field were asked to find these regions and cut the non-ROI parts out from the original picture with an image-editing program (GIMP).

2.2. Classification

The method used to assign polyps to one of the two groups (HP and non-HP) is based on normal human classification methods. A key aspect of them is to categorize the glands on the surface of polyps by their shapes. Based on Kudo-Tsuruta classification (see [5]), 4 categories were used: from the small, circular glands were called “sI”-type, to “sIV” category that stands for entirely irregular shape with filamentary structure. “sII” and “sIII” are transitional shapes, the former is for compact, but not circular shapes which indicates slightly distorted, the latter is for highly prolonged but not filamentary structure. Fig. 2 shows examples for each type.

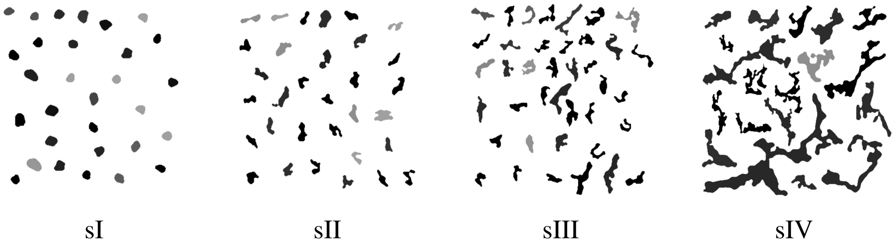


Figure 2. Examples for individual glands of different types

Based on prior researches sI and sII type shapes are indicators of HP, while sIII and sIV glands imply non-HP polyp. Hence the exposed method classifies the spots on the polyp image, calculates the relative area of sIII and sIV spots (with higher weight for sIV and sI spots). Higher value of this relative area indicates non-HP type.

To find a classification method for glands a lot of scale-invariant geometrical parameters were calculated. It turned out that the following three parameters can characterize these maculas rather accurately:

- Circularity: $C = 4\pi A/P^2$, where A is the area P is the perimeter of the spot.
- Solidity: $S = A/A_h$, where A_h is the area of the convex hull.
- Hu-0: H, the 0th Hu-invariant. (See [10])

To determine a spot type i.e to classify it we used QDA. As training set we extracted 1200 maculas from original colonoscopic images labeled by medical doctors who has huge experience on this field. The resulted QDA method can rank the shapes into the 4 sets with high precision (higher than 95%) based on C, S and H parameters.

An another indicator that correlates with the HP–non-HP groups is the width of the intensity histogram of polyp surface. Based on the previously mentioned "relative area" and this contrast-like parameter a classification method was developed. The statistical accuracy (85–90%) reached the accuracy of human classification with Nice and Kudo methods. (The reference in accuracy calculation was the result of histology.)

The details of these classification methods was published in [8].The statistical results can be found in [9]. Improved version will be published in a further paper.

3. The preprocessing

3.1. The aim of preprocessing step

Preprocessing in this classification software starts with noise filtering to reduce the impact of the camera's noise. After trying 3-4 different alternatives, median filter was selected as an appropriate step for doing this.

In endoscopic practice the medical doctor has limited freedom in the geometry and light sources, therefore significant illumination differences are perceptible in the images, even inside the ROI of one polyp. Fig. 3. shows three such examples.

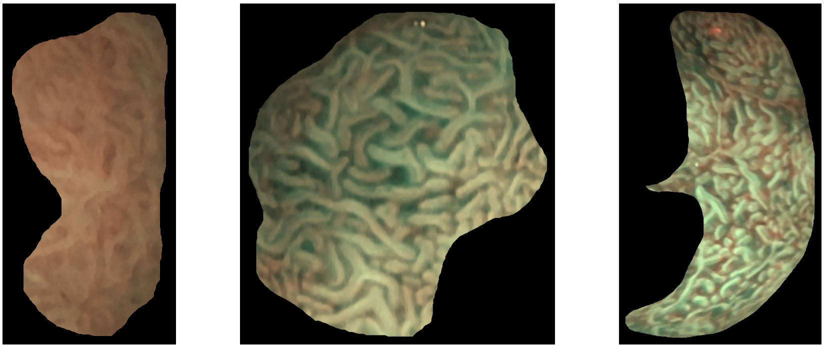


Figure 3. Examples of non-uniform illumination inside a polyp surface.

It is obvious that change of illumination makes it hard to detect the spots and measure their geometrical parameters accurately, and biases the histogram of the image. For this reason the main aim of preprocessing is to transform out the change of illumination.

Since surface of polyps are more or less ellipsoidal shapes, the characteristic size of illumination variation is in the order of the size of ROI. The standard high-pass filtering with Fourier transformation is not appropriate in this problem, because the ROI has irregular shape and it causes significant windowing bias.

3.2. Brightness correction step

Fig. 3 shows that illumination changes slowly but not monotonous inside the ROI. The basic idea of eliminating this effect is to approximate the illumination change with a "brightness function". Ratio of the image pixel values and this brightness function will describe the surface itself.

In absence of 3D data of polyp surface and light source geometry, a plausible approximation of this brightness function must be calculated from image data. The surface of the polyp has similar average properties across the whole ROI therefore the base assumption is to fit a function that slowly changes to the image intensity. Eventually this will produce a good approximation of the brightness function. After some consideration, third order, two variable polynomials was selected for this purpose.

To formulate the brightness correction step it is practical to use (x,y) coordinates in the image plane so that the image fit into the unit square. If the image has dimensions N_x and N_y, then the pixel in the i_x-th column and i_y-th row will have the following coordinates:

$$x = i_x \cdot \delta, \qquad y = i_y \cdot \delta, \qquad \text{where} \quad \delta = 1/\max(N_x,N_y). \tag{1}$$

The filtered image was converted to grayscale image with standard sRGB weights:

$$L[i_x,i_y] = 0.299R[i_x,i_y] + 0.587G[i_x,i_y] + 0.114B[i_x,i_y], \tag{2}$$

where $R[]$, $G[]$ and $B[]$ are the color components of the image and $L[]$ is the resulted grayscale image. (All color components are transformed into [0,1] interval.)

Let's define the following 3rd order 2 variable function in the plane of the image:

$$l(x,y) = A_{00}+A_{10}x+A_{01}y+A_{20}x^2+A_{11}xy+A_{02}y^2+A_{30}x^3+A_{21}x^2y+A_{12}xy^2+A_{03}y^3. \tag{3}$$

Method of least squares was used to find the A_{ij} parameters in this equation, when $l(x,y) = l(i_x \cdot \delta,i_y \cdot \delta)$ approximates $L[i_x,i_y]$. It means that the following function was minimized:

$$W(A_{00},A_{10},A_{01},\ldots,A_{03}) = \sum_{(i_x,i_y)\in ROI} (L[i_x,i_y] - l(i_x \cdot \delta,i_y \cdot \delta))^2. \tag{4}$$

(Note that "[]" means array indexes while "()" is for function arguments.)

Since $l(x,y)$ depends linearly on A_{kl} weights, extremal problem (4) can be solved explicitly with the traditional method of least squares. If we use stars for the optimal values (A^*_{00}, A^*_{10}, ...), the approximation of average brightness will be:

$$l^*(x,y) = A^*_{00} + A^*_{10}x + A^*01y + A^*_{20}x^2 + \ldots + A^*_{03}y^3. \tag{5}$$

Using this function, the brightness corrected, but not normalized image will be:

$$L_b[i_x,i_y] = \frac{L[i_x,i_y]}{\max(0.2,l^*(i_x\delta,i_y\delta))}. \tag{6}$$

Applying the "max" function in the denominator is for safety: in very rare occasions l^* has very low, even 0 or less value at some points. In this case it is not good for brightness correction.

The final step in this brightness correction step is the normalization:

$$L_{b,n}[i_x,i_y] = \frac{L_b[i_x,i_y]}{\max(L_b[i_x,i_y])}, \tag{7}$$

where the "max" is for the pixels of ROI.

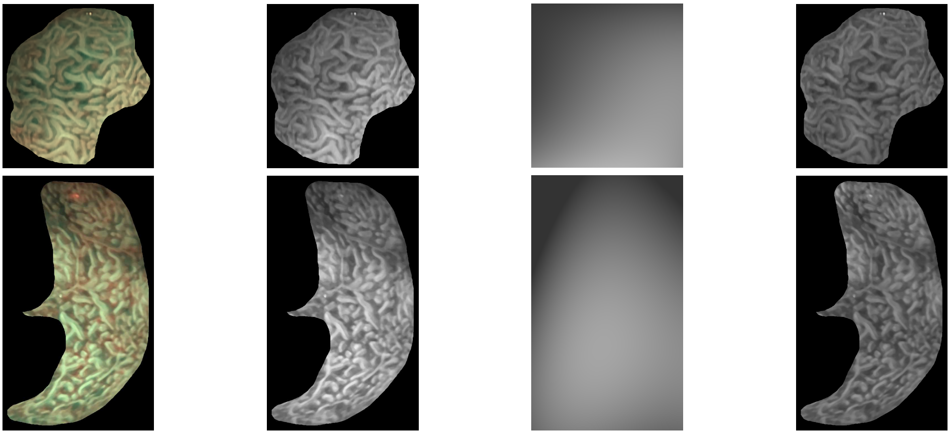

R, G, B L l^* $L_{b,n}$

Figure 4. Steps of brightness correction. See the text for annotations.

Fig. 4 shows two examples: original image, grayscale version, best-fit l^* and the final corrected and normalized $L_{b,n}$ image is displayed. It is clear that this type of brightness correction works well on arbitrary ROI, does not produce artifacts and reaches its aim: transforms out the effect of variable illumination.

3.3. Thresholding

On a grayscale image many detail is still hiding although most of these details are surplus. It is necessary to reduce the detailed intensity information into 1 bit: a pixel is inside of a "dark" spot or not. After this step we can measure the geometrical parameters of the spot.

The simplest method of this binarization is to use a fix threshold value for the whole image, i.e. the binary image will be:

$$B[i_x,i_y] = \begin{cases} 1, & \text{if} \quad L[i_x,i_y] > T \\ 0, & \text{otherwise} \end{cases} . \tag{8}$$

Selecting a global T threshold value is not self-explanatory. There are algorithms that can pick up a relatively good one, like the Otsu's method. ([11]) It provides an acceptable result if the average brightness of the picture is uniform, therefore it makes sense to apply it after the brightness correction.

3.3.1. Adaptive thresholding

However, choosing an adaptive threshold value is often makes the binarization more accurate. The usual way to do this is to calculate a place-dependent threshold value from a smoothed version of the image:

$$T[i_x,i_y] = \sum_{l=-k}^{k} \sum_{j=-k}^{k} L[i_x + j,i_y + l] \cdot K[j,l], \tag{9}$$

where K is the kernel of the smoothing operator, k is the "radius" of kernel. K is assumed to be normalized: $\sum_{l=-k}^{k} \sum_{j=-k}^{k} K[l,j] = 1$.

(9) means a simple convolution between L and K, therefore it can be rewritten as:

$$T = L * K, \tag{10}$$

where '*' is the sign of convolution.

The usual choices for K are: uniform and Gaussian.

- Uniform kernel: $K[j,l] = 1/(2k + 1)^2$ for all j and l.
- Gaussian kernel: $K[j,l] = n \cdot e^{-(j^2+l^2)/(2\sigma^2)}$ where σ is the deviation and n is a suitable normalization factor.

With that $T[]$ threshold array the binary image will be the following:

$$B[i_x,i_y] = \begin{cases} 1, & \text{if} \quad L[i_x,i_y] > T[i_x,i_y] + C \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

The offset value C is for fine-tuning: setting it to positive or negative value makes less or more 1-s in binary image.

The (9) formula is a convolution and can be calculated very efficiently with Fast Fourier Transformation for large image and kernel sizes. Therefore if we know the values of $L[]$, the kernel K and C offset value, binary image B can be determined in short time. This method is often referred as "adaptive thresholding".

3.3.2. Conditional adaptive thresholding

The original adaptive thresholding method has obvious problems if $L[]$ is not known for all the domain, and this is the case in endoscopic images of polyps, because of the irregular shape of ROI. Outside the ROI the pixels give no information about the polyp itself, therefore non-ROI pixels should be excluded from calculations. In other words: values of $L[]$ is unknown outside of the ROI.

A simple solution could be to use a constant value (e.g. 0) outside the ROI. However it is obvious that this would highly distort the result of convolution near the borders. A more sophisticated method is to use some "inpainting" method to fill the non-ROI pixels based on near-border values. This method was tested and rejected because a large area around the ROI should be filled with approximated values and the tests showed very unnatural intensity distributions. An additional drawback of inpainting is the high CPU-usage.

Instead of filling with constant or inpainted values an another approach was used. (9) can be interpreted as a weighted average of $L[]$ values (plus an offset). The sum of weights in K kernel is 1 in this case. It is easy to modify this formula to calculate the weighted average only for ROI pixels:

$$\begin{aligned} T[i_x,i_y] &= \frac{1}{D[i_x,i_y]} \sum_{(j,l)|(i_x+j,i_y+k)\in ROI} L[i_x+j,i_y+l]\cdot K[j,l] \qquad (12) \\ D[i_x,i_y] &= \sum_{(j,l)|(i_x+j,i_y+k)\in ROI} K[j,l]. \end{aligned}$$

The long formula below the sum sign means that the summation is for those (j,l) pairs, when (i_x+j,i_y+l) refers to a point inside the ROI.

It is obvious that (12) gives the same result as (9) for pixels deeply inside the ROI and always gives a normalized, weighted average of ROI pixels that are at most k distance in infinity norm. This kind of smoothing can work for arbitrary shaped ROIs, but in this form it is very expensive in computational steps. Fortunately it can be accelerated by reduction into two separate convolution. Let's define the characteristic function of the ROI:

$$Q[i_x,i_y] = \begin{cases} 1, & \text{if} \quad (i_x,i_y) \text{ is inside ROI,} \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

Using this function (12) can be written as:

$$\begin{aligned} T[i_x,i_y] &= \frac{1}{D[i_x,i_y]} \sum_{l=-k}^{k} \sum_{j=-k}^{k} (L[i_x+j,i_y+l]Q[i_x+j,i_y+l])K[j,l] \qquad (14) \\ D[i_x,i_y] &= \sum_{l=-k}^{k} \sum_{j=-k}^{k} Q[i_x+j,i_y+l]K[j,l]. \end{aligned}$$

On the right sides we have two simple convolutions:

$$\begin{aligned} T &= \frac{1}{D}(L \cdot Q) * K \qquad (15) \\ D &= Q * K. \end{aligned}$$

Both convolution can be effectively evaluated with FFT based algorithm. The pixelwise division with D array elements is a cheap operation but can occur division by zero operations. However, it is easy to show that it will never happen for points of ROI for all plausible K kernels. To be more precise: if all elements of K are nonnegative and $K[0,0] > 0$, then $D[i_x,i_y] > Q[i_x,i_y]K[0,0] = K[0,0] > 0$ for all (i_x,i_y) ∈ROI.

For non-ROI pixels, where D can be 0, T is arbitrary and the division with pixel of D is unnecessary, because values of T is used only in (11) to get the binarized image inside the ROI.

The steps in (15) is analogue with the conditional probability or conditional expected value calculations based on Bayes' theorem. The smoothing operator can be interpreted as a calculation method for a expected value of a pixel: the expected value depends on the values of neighbours through a weighted average calculation, where the close neighbours has higher weight.

4. Testing method

4.1. Test problems

The methods described above can be combined in many ways. For example, one can use Otsu's method for thresholding with or without brightness correction, or use adaptive thresholding with uniform or Gaussian kernel, or the last method: conditional adaptive thresholding, etc. To choose the best method, a set of test problems was constructed which are similar to the real polyp images but the list of dark spots and exact value of their geometrical parameters are known.

1. Four "flat" image were created using the categorized examples of dark spots. These were the same as the four image on Fig. 2. Each flat image contains spots of the same type which is useful to identify type-specific problems and inaccuracies.
2. Three different illumination distribution was created: a circular, an ellipsoidal and an ellipsoidal with flares. (See Fig. 5.)
3. The 12 test images were constructed by multiplication of flat images and illumination distributions with added Poisson-noise.

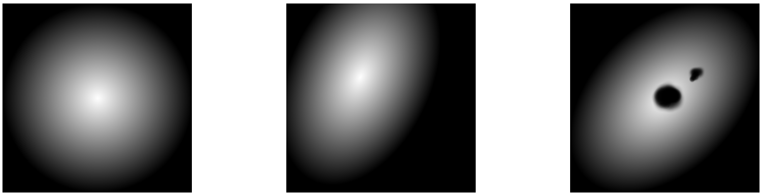

ID1 ID2 ID3

Figure 5. Illumination distributions of test problems.

The resulted 12 test images (combination of 4 flat and 3 illumination distribution images) has similar noise and brightness properties to real polyp images, but the exact geometrical parameters can be measured on the flat images. Two of them is displayed on Fig. 6. As a short notation "T-x/y" will refer to the test problem which was derived from x-th flat image and y-th illumination distribution.


T-1/3: sI*ID3+ noise T-4/3: sIV*ID3 +noise

Figure 6. Two of the 12 test images.

Note that the test images does not imitate the interesting parts of polyp images in every aspect. For example, on real images spots of all category can be found in one image, the spots are more randomly placed and sometimes are very close to each other. These simplified test images make it possible to measure the accuracy of preprocessing for every spot category and clear separation of spots prevent from unwanted artifacts.

On the illumination distribution images the pixel values are 0 outside the simulated polyps and inside the simulated flares, otherwise they are positive. Therefore the ROI of a test image is the area, where the illumination distribution pixels has positive values.

4.2. Testing method

The basic idea for testing is to measure the "exact" geometrical parameters on the original flat images (Fig. 2) and compare them to the preprocessed image values. However, some care is needed to take, because if a spot on a flat image is fully or partly outside of the ROI (nonzero pixels on Fig. 5) of a test image, one can not expect to find it after preprocessing. Therefore to get reference values of geometrical parameters, the flat images were masked by the ROI of test image and only these inner spots were used as reference.

The measured and stored parameters were the following: map of spots (the pixels for every individual spots), circularity (C), solidity (S), Hu 0 coefficient (H), area (A), perimeter (P).

The 12 test images were preprocessed for a number of combination of preprocessing steps and in each case the same parameters were measured on them.

Using the map of spots on the masked flat and the preprocessed test image, the matching spot-pairs was searched for at first. Two spots was called "matching" if they had at least one common pixel. For every spots on masked flat image the number of matching spots (M) on preprocessed image was calculated. In ideal case $M = 1$ for every spot, but if the preprocessing is inaccurate, we can get $M = 0$ (no matching spot) or $M > 1$ (multiple matching spots). The "miss ratio" of the spot-matching was the ratio of $M \neq 1$ cases and the number of spots. This "miss ratio" characterizes the accuracy of spot-preserving of a particular preprocessing method.

For spots with $M = 1$ matching pair, the geometrical parameters were measured also on preprocessed test image and the relative difference between corresponding values was calculated. The average of these relative errors was used to characterize the accuracy of geometrical parameter preservation.

4.3. The examined preprocessing methods

In section 3 a lot of possible preprocessing step was described. Some of them has parameters, and some of them are optional. This means that a lot of preprocessing chain can be formed using them. For a summary, see Table 1.

The number of possible preprocessing methods is very high, so all of the test results can

Preproc. step	**Variation**	**Parameters**	**Abbrev.**
Median filter	–	radius (R)	–
Brightness corr.	–	–	BC
Thresholding	Otsu	–	OT
	Adaptive / Uniform kernel	kernel radius (k)	AThu
	Adaptive / Gaussian kernel	kernel radius (k)	AThg
	Conditional Adaptive / Gaussian	kernel radius (k)	CATh

Table 1. Preprocessing steps used in testing.

not be presented in this paper. The calculations showed that the best parameters are the following:

- Median filter radius: $R = 1$ (3x3 size structuring element).
  $R = 0$ means no median filtering. In this case the noise is not suppressed what leads to big errors. $R > 1$ means strong noise reduction but it smooths the boundary of spots and results in high error level in geometrical parameters.
- Adaptive threshold kernel radius: $k = W/3$, where W is an effective width of the ROI calculated as the ratio of area and diameter of ROI.
  A lot of test calculations showed that the kernel radius must be in the magnitude of average width of ROI. Calculations was performed with different values (i.e. $k = W$, $W/2$, $W/3$, $W/5$ ...) and the $k = W/3$ choice appeared to be the best.

4.4. Implementation

The described methods were implemented in Python and NumPy module was used for numerical calculations (Brightness correction, CATh). For basic image manipulation, input-output, convolution, standard adaptive thresholding (AThu, AThg) OpenCV was applied. ([12])

5. Results

The relative errors of geometrical parameters are denoted with $\mathcal{R}(C)$, $\mathcal{R}(S)$, ..., the average of them is $\overline{\mathcal{R}}$. The number of spots on masked flat image is N, and N_1, N_0, N_m stands for number of unique match, missing spot and multiple matches respectively. The miss ratio is $\mathcal{M} = (N_0 + N_m)/N$.

A specific preprocessing method is characterized by this average relative error ($\overline{\mathcal{R}}$) and miss ratio ($\mathcal{M}$). In ideal case both one would be 0.

5.1. A detailed calculation result

Tab. 2 shows the results for T-1/2, T-2/3, T-3/3, T-4/3 test results (the cases for the most difficult ID3 illumination distribution), and for $M = 1$ (i.e. for 3x3 median filter). All of the measured parameters (relative errors of all 5 geometrical parameters and all type of matching errors) is displayed in details. The results are sorted in increasing order of $\mathcal{M}$ inside a test problem and increasing $\overline{\mathcal{R}}$ value if the miss rate is the same.

Method		Geometrical Errors						Matching Errors				
BC	Thresh	$\mathcal{R}(C)$	$\mathcal{R}(S)$	$\mathcal{R}(H)$	$\mathcal{R}(A)$	$\mathcal{R}(P)$	$\overline{\mathcal{R}}$	N	N_1	N_m	N_0	$\mathcal{M}$
T-1/3												
+	CATh	0.072	0.017	0.003	0.024	0.043	3.18%	20	20	0	0	0.00%
+	AThg	0.082	0.025	0.003	0.014	0.044	3.36%	20	20	0	0	0.00%
-	CATh	0.075	0.019	0.003	0.026	0.046	3.40%	20	20	0	0	0.00%
-	AThg	0.086	0.028	0.003	0.016	0.047	3.59%	20	20	0	0	0.00%
+	AThu	0.108	0.038	0.015	0.058	0.060	5.59%	20	20	0	0	0.00%
-	AThu	0.170	0.112	0.212	0.401	0.717	32.25%	20	19	0	1	5.00%
-	OT	0.212	0.187	0.047	6.941	5.389	255.51%	20	17	0	3	15.00%
+	OT	0.046	0.012	0.022	0.162	0.110	7.05%	20	15	0	5	25.00%
T-2/3												
+	CATh	0.077	0.021	0.014	0.028	0.047	3.75%	24	24	0	0	0.00%
-	CATh	0.075	0.019	0.016	0.031	0.049	3.82%	24	24	0	0	0.00%
-	AThg	0.100	0.041	0.010	0.020	0.055	4.53%	24	23	1	0	4.17%
+	AThg	0.096	0.038	0.011	0.026	0.057	4.56%	24	23	1	0	4.17%
+	AThu	0.160	0.071	0.019	0.088	0.096	8.69%	24	22	2	0	8.33%
-	AThu	0.195	0.132	0.193	1.329	0.988	56.75%	24	22	2	0	8.33%
-	OT	0.262	0.242	0.278	4.959	2.765	170.12%	24	20	0	4	16.67%
+	OT	0.114	0.087	0.074	0.224	0.150	12.97%	24	16	0	8	33.33%
T-3/3												
+	CATh	0.087	0.064	0.010	0.041	0.057	5.17%	29	28	1	0	3.45%
-	CATh	0.083	0.065	0.010	0.046	0.058	5.23%	29	28	1	0	3.45%
-	AThg	0.126	0.049	0.021	0.049	0.078	6.44%	29	28	1	0	3.45%
+	AThg	0.176	0.118	0.033	0.074	0.103	10.09%	29	28	1	0	3.45%
-	AThu	0.210	0.148	0.149	0.626	0.631	35.27%	29	27	2	0	6.90%
+	AThu	0.155	0.097	0.025	0.086	0.103	9.31%	29	26	2	1	10.34%
-	OT	0.342	0.363	0.069	12.266	6.282	386.43%	29	22	0	7	24.14%
+	OT	0.080	0.094	0.058	0.216	0.140	11.76%	29	19	1	9	34.48%
T-4/3												
+	CATh	0.070	0.057	0.012	0.034	0.047	4.40%	19	19	0	0	0.00%
-	CATh	0.067	0.058	0.012	0.037	0.047	4.43%	19	19	0	0	0.00%
+	AThg	0.461	0.220	0.057	0.120	0.136	19.87%	19	16	3	0	15.79%
+	AThu	0.450	0.224	0.056	0.132	0.139	20.01%	19	16	3	0	15.79%
-	AThg	0.489	0.229	0.058	0.125	0.142	20.88%	19	15	4	0	21.05%
-	AThu	0.355	0.162	0.033	0.127	0.119	15.92%	19	14	3	2	26.32%
-	OT	0.443	0.378	0.143	8.671	3.480	262.31%	19	13	2	4	31.58%
+	OT	0.263	0.151	0.115	0.232	0.187	18.94%	19	11	2	6	42.11%

Table 2. Detailed test results with $R = 1$ median filtering for the ID3 illumination distribution for all the flat images.

The most important results can be summarized as follows:

- The Otsu method is just not good enough even with brightness correction. It is in accordance with our preliminary suspicion. Although it should be noted that with brightness correction the geometrical error is quite low, but the miss rate is still unacceptably high.

- The classical adaptive thresholding works significantly better with Gaussian kernel compared to uniform one. It is plausible: the uniform kernel uses the same weight factor for far and near pixels.
- The clear winner is the CATh (Conditional Adaptive Thresholding) method with BC (Brightness correction). CATh was surprisingly accurate even without BC step.
- CATh works well in the following manner too: flat images with higher index contain more complex and elongated shapes. The miss rate and average geometrical error remains almost the same with this method for T-4/3 test case (sIV spots) than that of with T-1/3 case (sI spots).

The detailed error quantities ($\mathcal{R}(C)$, $\mathcal{R}(S)$, ..., N_m, N_0) show no interesting results: they are in good correlation with the total parameters ($\mathcal{R}$ and $\mathcal{M}$). Therefore in the following test results only $\mathcal{R}$ and $\mathcal{M}$ will be displayed.

To keep this paper compact, results for ID1 and ID2 illumination distribution is not presented. As they are less complex than the ID3 case, both $\mathcal{R}$ and $\mathcal{M}$ has smaller values than in ID3 cases, but the relative accuracy of preprocessing methods is the same as above.

5.2. The effect of the median filter's radius

The ideal radius of median filter (R) was also examined. Higher values mean more noise reduction but can significantly change the shape of the spots. In this paper only the results for the best preprocessing method (CATh with BC) is presented for ID3 illumination distribution: Table 3 shows the miss rates, Table 4 the average geometrical errors.

	$R = 1$	$R = 2$	$R = 3$	$R = 7$
T-1/3	0.00%	0.00%	0.00%	0.00%
T-2/3	0.00%	0.00%	0.00%	0.00%
T-3/3	3.45%	3.45%	0.00%	3.45%
T-4/3	5.26%	0.00%	0.00%	0.00%

Table 3. Missing rates ($\mathcal{M}$) for the best preprocessing method (CATh with BC) with different radius of median filter (R).

It is clear that $R = 1$, which is equivalent with no median filtering is the worst in miss rates. The matching errors arise in the sIII and sIV cases, i.e. when the shapes are more complex. The reason is that in this cases the noise can lead to division of one spot into smaller ones.

	$R = 1$	$R = 2$	$R = 3$	$R = 7$
T-1/3	1.83%	3.18%	3.94%	5.71%
T-2/3	2.55%	3.75%	4.94%	10.78%
T-3/3	4.14%	5.17%	6.15%	13.78%
T-4/3	3.90%	4.40%	3.80%	10.40%

Table 4. Average geometrical errors ($\overline{\mathcal{R}}$) for the best preprocessing method (CATh with BC) with different radius of median filter (R).

With increasing median filter radius the average geometrical error is increases in most of the cases. (There is only one exception: T-4/3 is better with $R = 3$ than with $R = 2$.)

One can conclude that median filtering is necessary and the optimal value of its radius is 2 or 3. This is the reason why we displayed $R = 2$ cases above.

Note that similar calculations was performed with other noise filtering than median filter, but they could not produce better results and the most of them required parameter-tuning.

5.3. Test calculation without noise

As it was shown above, presence of noise makes filtering necessary, but the filtering has negative effect on geometrical parameters. To measure the accuracy of preprocessing methods without the bias of noise, a new text calculation was performed with test images without simulated noise. Naturally, the median filtering was abandoned during the preprocessing method for these images. The results of this noiseless calculations is displayed in Table 5. (For comparison, results with noise and $R = 2$ median filtering is printed there.)

	with noise ($R = 2$)	without noise
T-1/3	3.18%	1.65%
T-2/3	3.75%	2.54%
T-3/3	5.17%	4.87%
T-4/3	4.40%	3.73%

Table 5. Average geometrical errors ($\overline{\mathcal{R}}$) for the CATh+BC method: calculation with and without simulated noise.

These average relative error values show the accuracy of the BC+CATh method itself, while the difference between two columns shows the effect of the noise. We can conclude that the noise increased the geometrical errors only with approximately 1% and the own error of BC+CATh method is in the order of 2–5%, which is adequately low for our purposes.

6. Conclusion

The results displayed above show that combination of brightness correction (BC) and conditional adaptive thresholding (CATh) steps is an appropriate method for preprocessing colonoscopic images of polyps. This method preserves the spots and their geometrical parameters more accurately than classical adaptive thresholding ones. The described methods also provide a firm base for further research in the field of automatic polyp classification, but CATh seems to be useful in other areas, where smoothing or adaptive thresholding calculations are needed on an irregular region of interest.

References

[1] Neumann H, Vieth M, Fry LC, Günther C, Atreya R, Neurath MF, Mönkemüller K: Learning curve of virtual chromoendoscopy for the prediction of hyperplastic and adenomatous colorectal lesions: a prospective 2-center study; Gastrointestinal Endoscopy, Vol. 78, No. 1, pp 444-449, 2013
DOI:10.1016/j.gie.2013.02.001

[2] Rex DK: Colonoscopy. Gastrointestinal Endoscopy, Vol. 78, No. 3, 2013
DOI:10.1016/j.gie.2013.06.025

[3] Fiori M, Musé P, Sapiro G: A Complete System for Candidate Polyps Detection in Virtual Colonoscopy International Journal of Pattern Recognition and Artificial Intelligence Vol. 28, No. 7, 2014
DOI: 10.1142/S0218001414600143

[4] Tomonori Yano, Hironori Yamamoto, Keijiro Sunada, Tomohiko Miyata, Michiko Iwamoto, Yoshikazu Hayashi, Masayuki Arashiro, Kentaro Sugano: Endoscopic classification of vascular lesions of the small intensitine; Gastrointestinal Endoscopy, Vol. 67, No. 1, pp. 196-172, 2008
DOI:10.1016/j.gie.2007.08.005

[5] Yoshito Takemura, Shigeto Yoshida, Shinji Tanaka, Keiichi Onji, Shiro Oka, Toru Tamaki, Kazufumi Kaneda, Masaharu Yoshihara, Kazuaki Chayama: Quantitative analysis and development of a computer-aided system for identification of regular pit patterns of colorectal lesions; Gastrointestinal Endoscopy, Vol. 72, No. 5, pp.1047-1051, 2010
DOI:10.1016/j.gie.2010.07.037

[6] Arber N, Spicak J, Racz I, Zavoral M, Breazna A, Gerletti P, Lechuga MJ, Collins N, Rosenstein RB, Eagle CJ, Levin B: Five-year analysis of the prevention of colorectal sporadic adenomatous polyps trial, American Journal Of Gastroenterology Vol. 106, No. 6, pp. 1135-1146. : 2011
DOI:10.1038/ajg.2011.116

[7] Durcsán H, Szabó A, Kovács V, Szalai M, Kiss Gy, Lakó K, Gyulyás J, Regőczi H, Rácz I: Population based Colorectal Cancer Screening; preliminary results from Győr, Zeitschrift Für Gastroenterologie 5, pp. 53, 2015

[8] Horváth A, Spindler Sz, Rácz I, Szalai M: Kolonoszkópiás polipképek számítógépes kategorizálása; In proceedings of: "Képfeldolgozók és Alakfelismerők Társaságának 10. országos konferenciája." Kecskemét, Hungary, 2015.01.27-30., pp. 505-514, 2015

[9] Rácz I, Horváth A, Szalai M, Spindler S, Kiss G, Regöczi H, Horváth Z: Digital Image Processing Software for Predicting the Histology of Small Colorectal Polyps by Using Narrow-Band Imaging Magnifying Colonoscopy. Gastrointestinal Endoscopy, Vol. 81, No. 5, AB259, 2015

[10] Hu, Ming-Kuei: Visual pattern recognition by moment invariants. Information Theory, IRE Transactions on 8.2, pp. 179-187: 1962
DOI 10.1109/TIT.1962.1057692

[11] Otsu, Nobuyuki: A threshold selection method from gray-level histograms. Automatica 11, no. 285-296 pp. 23-27, 1975

[12] Bradski G, Adrian K: Learning OpenCV: Computer vision with the OpenCV library. O'Reilly Media, Inc., 2008

Acta
Technica
Jaurinensis

Vol. 9, No. 1, pp. 83-96, 2016
DOI: 10.14513/actatechjaur.v9.n1.400
Available online at acta.sze.hu

Determination of Load for Quasi-static Calculations of Railway Track Stress-strain State

D. Kurhan

The Dnepropetrovsk National University of Railway Transport named after academician V. Lazaryan, Department of Railway and Railway's Facilities
Lazaryan st. 2, 49010, Dnepropetrovsk, Ukraine
Phone: +38 056 373 15 42
e-mail: kurgan@brailsys.com

Abstract: When calculating the railway track stress-strain state one usually assumes that total strains are brought immediately from applied load and the process dynamics is taken into account by the respective levels of design force. The dynamic component of the design force depends on various factors that are not always taken into account to the full. The analytical analysis of the calculation methods and the experiment testing data resulted in the following recommendations: for freight trains, especially in the conditions of soft rail support, it is advisable to take into account the effect of adjacent wheels; for modern passenger cars there is no significant load dependence on speed, and the main factor of dynamic component is the track fluctuations.

Keywords: railway, permanent way, dynamic load, track stress-strain state

1. Introduction

Depending on the problem to be solved, one can use both relatively simple two-dimensional design models and developed models, which are described with the systems including dozens of equations. Despite the fact that this refers to the interaction of track and rolling stock, still the problems focused on the rolling stock study, and those focused on the railway track study have fundamental differences. Rolling stock models are, in most cases, the systems of motion (vibrations) of the interconnected solids. Typically, such models are mathematically described by Lagrange equations. Railway track operation is more naturally described not as motion of solids, but the strains thereof. Therefore, the railway track is more often mathematically described by the models based directly on theory of elasticity or its numerical representations in the form of finite element method and others. The combination of different approaches in one model greatly complicates their creation and subsequent application so is impractical for most tasks.

Thus, the railway track models in most cases come to a system of bodies (or layers) with elastic strain under load. These models, as a rule, are quasi-static: one of the main conditions at their creation is that the strain caused by the applied load are brought immediately and in full constant volume (as during static load), and the process dynamics is calculated by the respective levels of the applied load, with addition of the necessary complements to its static value.

Today, there are various methods allowing determination of track external load design values. The purpose of this work is to analyse these methods, to study the results of experimental testing of modern rolling stock action on track and to provide recommendations for their use.

2. Methods for determining the design force acting on the rail

2.1. Probabilistic approach using statistical population of dynamic complements

A probabilistic approach using statistical population of dynamic complements is the base one to determine the wheel design force acting on the rail in order to perform the so-called “Track strength and stability calculations”. This calculation is the official method in Ukraine [1] and some other countries. It is used to solve such problems as determination of stresses and strains in the railway track elements caused by the rolling stock impact, determination of the required strength of permanent way for the set operation conditions, determination of operation conditions for the set track design (including allowable speeds in terms of strength and temperature mode of continuous welded rail operation), etc.

The calculation basis is the hypothesis that the wheel force acting on the rail has probabilistic nature and is subject to the law of Gaussian distribution. The example of the respective distribution curve is presented in the work [2], is shown in Fig. 1. The result is referred to the rail stress values observed for the same wheel load at the same speed.

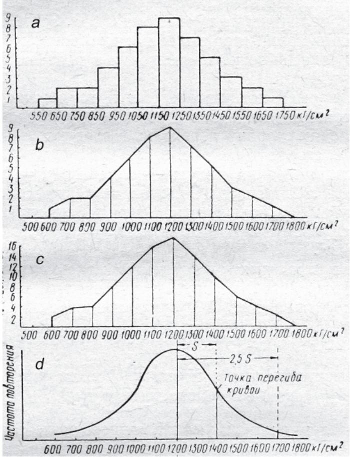


Figure 1. Probability of rail stress distribution [2]: a – distribution histogram; b – distribution polygon; c – frequency polygon; d – Gaussian distribution curve

Most often there is a mean wheel force acting on the rail, but there can be both larger and smaller values. It is assumed that the design force is determined by the formula [1…3]:

$$Q_{\text{dyn}} = \bar{Q} + \lambda S\,, \tag{1}$$

where

Q_{dyn} : calculated (dynamic) value of force;

$\bar{Q}$: mean value of force;

λ : probability factor, is taken λ =2.5, corresponding to the non-exceedance probability of design force 0.994;

S : root-mean-square deviation of force.

Mean value of dynamic force is made up of static load (vehicle weight referred to one wheel (Q_{stat}) and the sum of mean values of dynamic complements ($\bar{Q}_i$):

$$\bar{Q} = Q_{\text{stat}} + \sum \bar{Q}_i\ . \tag{2}$$

Total root mean square deviation consists of the geometric sum of the root-mean-square deviations of dynamic complements (S_i):

$$S = \sqrt{\sum S_i^2}\ . \tag{3}$$

Thus, each dynamic component is presented as the composition of its mean and root-mean-square deviation. As dynamic complements the following forces are taken into account: additional force due to vehicle bolster structure vibrations, inertial force due to wheel movement on the track bumps, inertial force of unsprung part from isolated irregularities on the wheel and inertial force of unsprung part from continuous irregularities on the wheel. The method of calculation of these dynamic complements is a set of analytical equations and approximations of experimental data. It is presented in the works [1…3] and others.

Thus, the considered approach makes it possible to determine the design force value taking into account the speed of movement and some key parameters associated with the design and the condition of railway track and rolling stock.

2.2. Determination of dynamic force through static load

The dynamic load of the track can be calculated from the static loads [4]:

$$Q_{\text{dyn}} = Q_{\text{stat}} + t \cdot \bar{s} \cdot Q_{\text{stat}}\,, \tag{4}$$

$$\bar{s} = n \cdot \varphi\,, \tag{5}$$

$$\varphi = 1 + \frac{V - 60}{140}, \tag{6}$$

where

t : distribution factor, if t =3 the accurary of calculation is 99.7%;

n : 0.1…0.3 (depends on the condition of trac);

φ : speed factor (for speeds up to 60 km/h φ =1 [5]);

V : speed in km/h.

The equation (6) may have a different look, for example, to take into account the type of the train (passenger or cargo) [5].

Examples of similar calculations: model for determination of the stress distribution in the longitudinal direction of the track [6]; the dynamic train loading was converted into an equivalent creep stress, using an equivalent static force method [7]; the investigation was aimed at static and dynamic load rating of aged railway concrete sleepers after service [8].

2.3. Calculation of adjacent wheels action

Wheels, located next to the calculated one, especially within the bogie, can affect the load level on the track. According to [1] such effect must be taken into account.

The basis is the known differential equation for rail deflection on equielastic support, used in many works, for example [2, 3, 9]:

$$\frac{d^4 z}{dx^4} + \frac{U}{EI} z = 0, \tag{7}$$

where

z : vertical rail deflection;

x : distance on rail from the force application point;

U : modulus of rail support elasticity in the vertical plane;

E : modulus of rail steel elasticity;

I : moment of rail inertia.

Solution of the equation (7) will be the rail lengthwise deflection function

$$z(x) = \frac{Qk}{2U} e^{-kx} (\cos kx + \sin kx), \tag{8}$$

$$k = \sqrt[4]{\frac{U}{4EI}}, \tag{9}$$

where

Q : load taken for calculation

Reverse solving of the equation (7) allows finding the effect of the force at a distance “x” on the track design section. The method of “Track strength and stability calculations” [1] recommends transition from the design force to the equivalent one. The action of such a force must be equivalent to the load caused by the combination of several wheels. Given that the stress dependence on the distance to the force application point for rails and rail support is different, one determines two equivalent forces. The first equivalent force ($Q_{\text{ekv}}^{\text{I}}$) is used to determine the bending moment and the stress in the rail, the second one ($Q_{\text{ekv}}^{\text{II}}$) – to determine the rail deflection and the stress in the rail support elements (in sleepers, ballast, roadbed):

$$Q_{\text{ekv}}^{\text{I}} = Q_{\text{dyn}} + \sum \bar{Q}\mu_i , \tag{10}$$

$$Q_{\text{ekv}}^{\text{II}} = P_{\text{dyn}} + \sum \bar{Q}\eta_i , \tag{11}$$

$$\mu_i = e^{-kx_i}\left(\cos kx_i - \sin kx_i\right), \tag{12}$$

$$\eta_i = e^{-kx_i}\left(\cos kx_i + \sin kx_i\right), \tag{13}$$

The formulas (10) and (11) are presented in the form meeting the hypothesis that the calculated wheel (which coincides with the calculated track section) transmits the calculated (dynamic) value of the force, and all other wheels – the mean value. Example of the variant, when the calculated wheel is the middle wheel of a three-axle bogie, is shown in Fig. 2 [3].

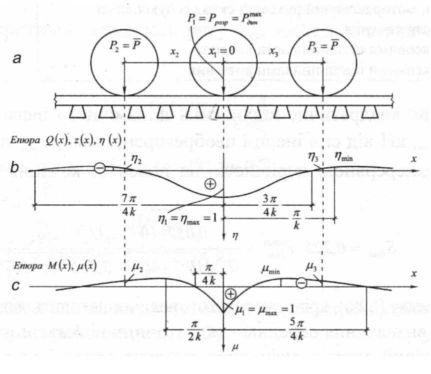


Figure 2. Load from three-axle bogie [4]: a – diagram of three-axle bogie; b – factor influence line η(x); c – factor influence line μ(x)

3. Experimental research of modern passenger train effect on the track

The purpose of experimental research was practical assessment of dynamic effect on the track caused by the modern passenger trains to be operated in Ukraine with a fairly high speed. The experimental trials covered Talgo and Skoda trains. Talgo train consisted of KZ4A locomotive and articulated passenger cars adjacently rested on uniaxial bogies. Skoda train consisted of motor and towed cars separately rested on two-axle bogies. The diagrams of the experimental trains are shown in Fig. 3 and 4.

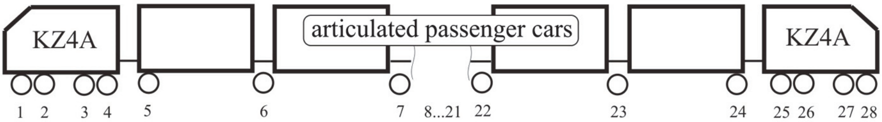


Figure 3. Diagram of experimental Talgo train

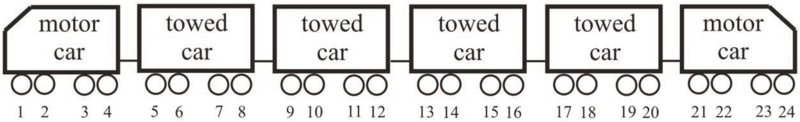


Figure 4. Diagram of experimental Skoda train

Experimental tests occurred on straight sections of the track without swerves requiring speed limit. The permanent way of both experimental sections was consisted of continuous welded rail R65, concrete sleepers, crushed stone ballast (including sub-ballast) of min 60 cm below the sleepers. Maximum speed was 176 and 200 km/h for Skoda and Talgo trains accordingly.

Organization and experiments were conducted by "Railway Track Testing Laboratory" of the Dnepropetrovsk National University of Railway Transport.

Some of the main indicators measured during the experimental train passing along the tested section were the rail stresses in several sections on top, web and base. Example of strain-gauge transducer installation in the rail section is shown in Fig. 5, example of stress record from the software window, which processed the data [10], – in Fig. 6.

The value of wheel vertical force acting on the rail was calculated based on the rail stress measurement results:

$$Q = 4kW\sigma\,, \tag{14}$$

where

W : rail moment resistance;

σ : semi-sum of stresses, measured on the outside and the inside edge of the rail base.

Figure 5. Installing transducers on a rail

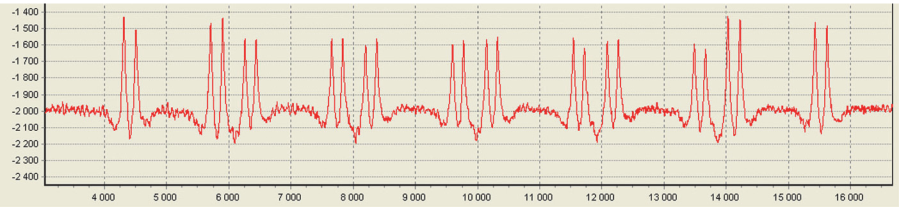


Figure 6. Example of rail stress record (stress on the base edge due to Skoda train passing at a speed of 176 km/h)

The results of rolling stock unit weighing allowed receiving the static load level. Processing of statistic data showed sufficient compliance of resulting distribution of vertical force values with Gauss' law, Fig. 7. For each speed level we determined the mean value and the root-mean-square deviation of dynamic force. Some tasks of strain accumulation process modelling and other demand not only the most probabilistic value of the force, but also a range of its possible values. And to assess the wheel stability it is appropriate to conduct the calculations taking into account the probabilistic minimum value of vertical force. The design dynamic force was determined as the maximum (minimum) one with increase (decrease) probability of 0.994:

$$Q_{\mathrm{dyn}} = \overline{Q} \pm 2.5S\,. \tag{15}$$

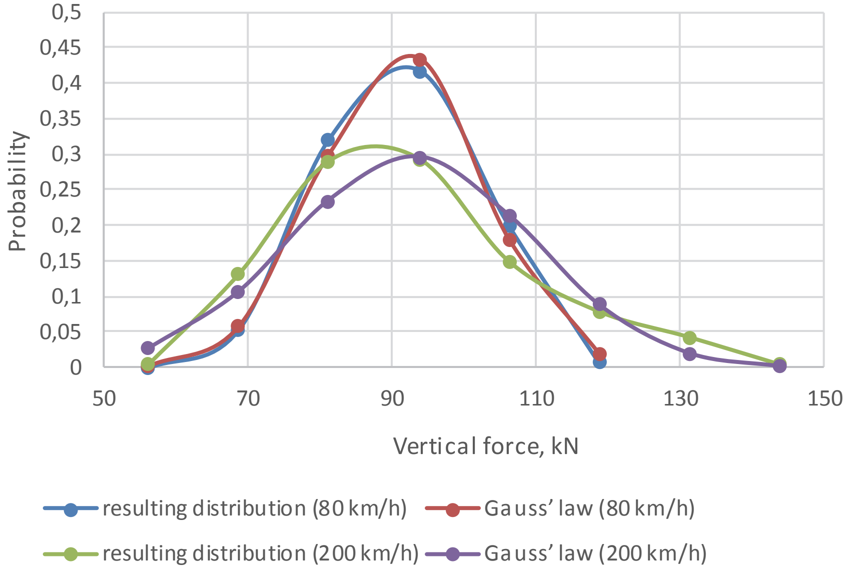


Figure 7. Law of distribution of the wheel vertical force on rail for Talgo train

The data processing results are shown in Fig. 8…11. For visual analysis, they are presented in the same horizontal and vertical scale.

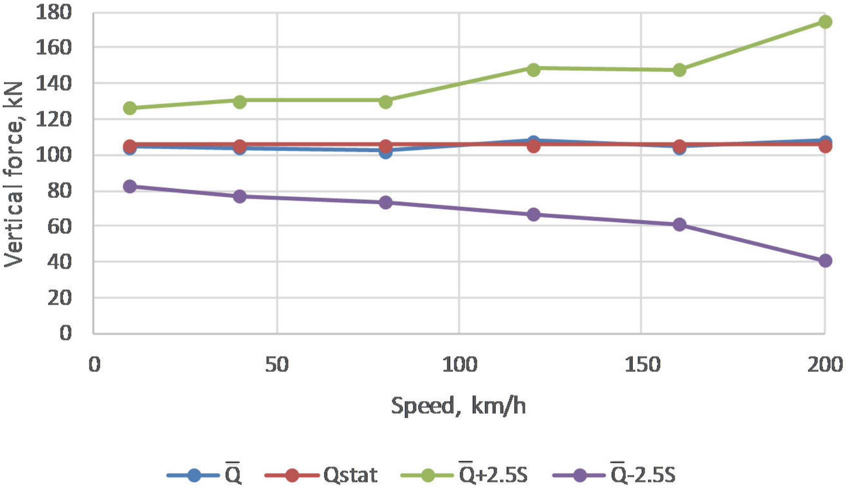


Figure 8. Vertical forces by experimental research on Talgo passenger train (for KZ4A locomotive)

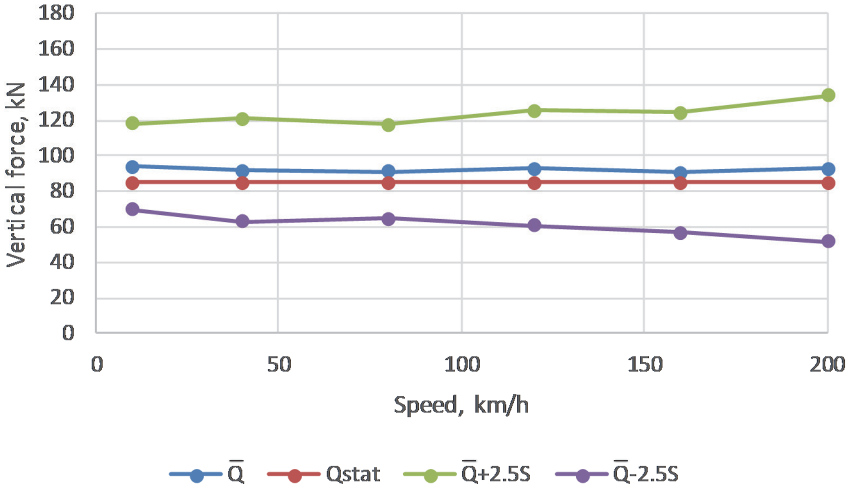


Figure 9. Vertical forces by experimental research on Talgo passenger train (for a car)

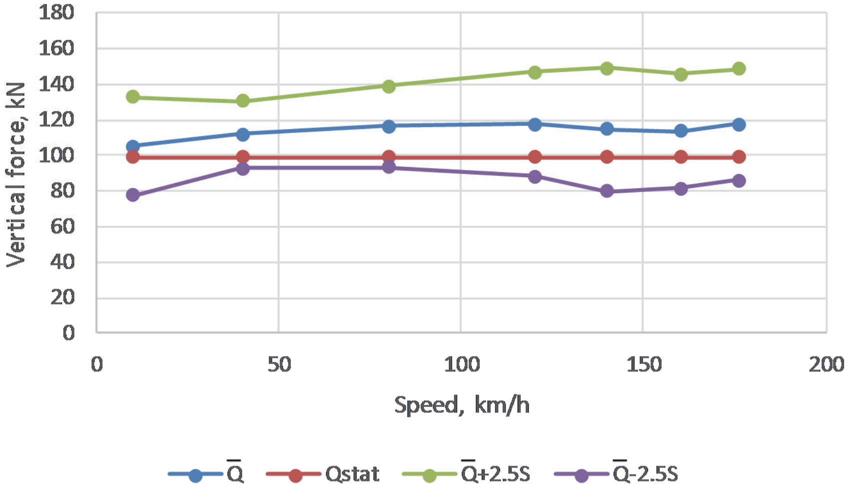


Figure 10. Vertical forces by experimental research on Skoda passenger train (for a motor car)

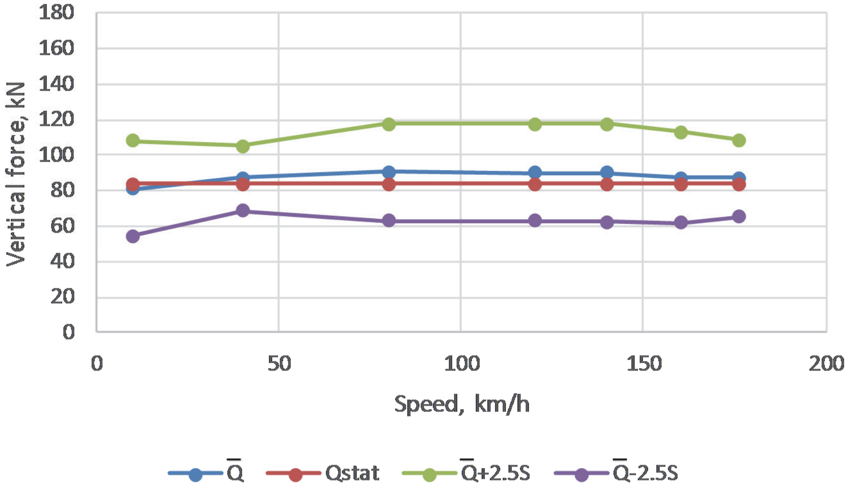


Figure 11. Vertical forces by experimental research on Skoda passenger train (for a motorless car)

4. Analysis of factors affecting the value of dynamic force

4.1. Effect of bogie design and rail support elasticity

In most cases the modern passenger and freight cars have two-axle bogies, and the locomotives have three or two-axle ones. Some methods of determining the rolling stock force acting on the track require taking into account of the bogie second (third) wheel effect, while the others ignore it.

The analysis of such effect was performed based on the formulas (10…13). The result depends mainly on two factors: distance from the calculated track section to the wheel and elastic modulus of the rail support. Particular attention was paid to the following groups of rolling stock: freight cars on CNII-H3-0 bogies with 185 cm axle base; locomotives with 210 cm axle base (such as 2TE10, M62 and others); passenger cars on KVZ-CNNI bogies with 240 cm axle base. As a rule, the locomotives with small axle base have three-axle bogies; in this case the track section under the middle axis is subject to the double additional load caused by outermost wheels.

The average values of the wheel effect on the track depending on the distance are shown in Fig. 12. As it can be seen, the pressure on the rail (correspondingly the rail section bending moment and the rail stress) away from the calculated section decreases rapidly and even enters the unloading area. Thus, when solving the rail stress tasks, the adjacent wheel effect can be neglected. However, the pressure on rail support (accordingly the rail deflection, the stress in sleeper, ballast, roadbed) may increase by more than 5% for freight cars and locomotives with a small rigid wheelbase (for the latter ones the results shown in Fig. 12 may be doubled for a three-axle bogie). For areas with low modulus of rail support elasticity the pressure increase on rail support may be significant, for example, up to 15% at 20 MPa modulus of elasticity.

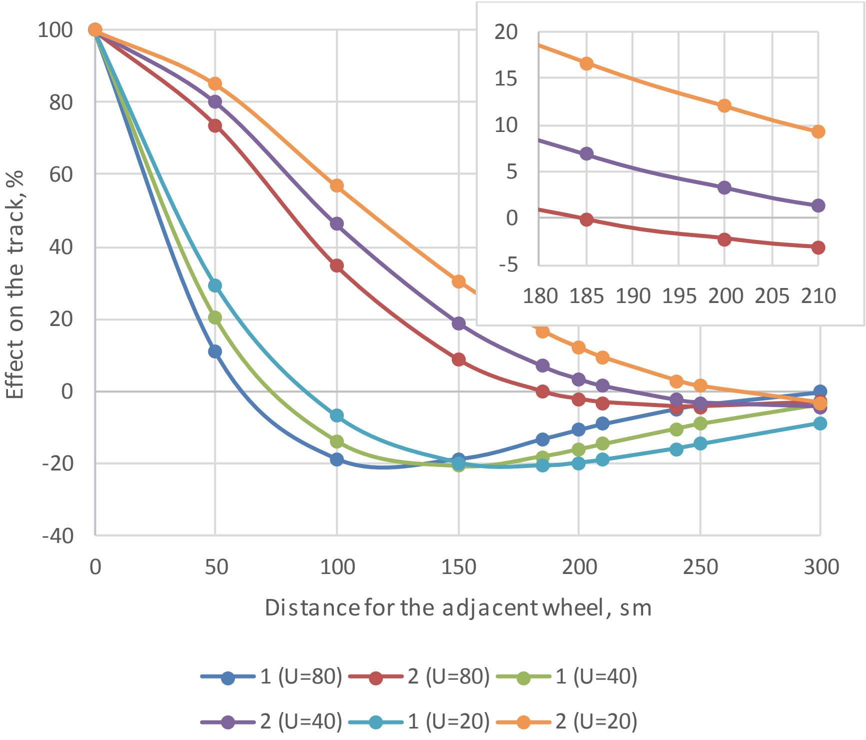


Figure 12. Wheel effect on the track depending on the distance for the modulus of rail support elasticity of 20, 40 and 80 MPa: 1 – pressure on the rail; 2 – pressure on the rail support

The problem of operating the railway line built on soft soils with a small modulus of rail support elasticity is relevant for many European countries. Various scientific works investigated the dynamic effects that occur when driving on the track with such characteristics, especially in case of sufficient axial loading (freight trains) or high speeds (passenger trains) – [11…15]. Such operating conditions require the relevant rail support stabilization [16, 17], justified by experimental research and mathematical modelling [18] with the corresponding characteristics of railway track and the rolling stock loading.

4.2. Effect of speed

Modern passenger cars have significantly improved dynamic performance, primarily due to the transition from mechanical to hydraulic spring systems with automated control. This approach prevents from the dynamic load growth caused by speed increase. This conclusion is supported by experimental research conducted by the author. Thus, for the locomotive we observe almost linear dependence of maximum probabilistic force on the speed (Fig. 8) that corresponds to calculations according to the existing methods [1, 4]. For motorless cars such force remains without significant changes even for high speeds (Fig. 9, 11). Similar conclusions for other types of rolling stock were obtained in the works [19, 20].

Even without expressed speed-dependence the vertical force dynamic component for passenger cars is evident. According to [1, 2, 3] the main factors of the dynamic component are vehicle bolster structure vibrations and track vibrations due to its elastic properties and geometric irregularities.

Accordingly, to the experimental rail stress measurements for Talgo train was conducted factor variance analysis in order to obtain the numerical characteristics of various factors affecting the value of wheel vertical force acting on rail. Thus, the effect of the axle number was analysed (examined 8 axles of cars with approximately equal static load), which describes the effect of the structure and the condition of rolling stock. Also the work analysed the influence of rail section point (examined 8 sections with different distances on the area without sufficient swerves), which describes the effect of the wheel motion on track dynamic bumps that arise from rail vibrations. The total number of measurements in the observability matrix varied for different speed levels. The smallest number (for speed of 200 km/h) amounted to 560 values. For different speed values the assessment results of the considered factors effect do not change fundamentally. For the range of 40…200 km/h we obtained the effect degree (by F-test) of the car axle number at 1.7; the effect degree of the rail section number –120.0 while the level of F-test critical value for considered samples equals 2.02.

The performed statistical analysis confirms numerically that the body vibrations in modern passenger cars are efficiently damped and do not lead to a significant increase in wheel vertical pressure force on rail. The main dynamic force perturbation factor can be considered as the wheel passing over the dynamic track bumps, which appears even in the absence of significant geometrical irregularities due to the rail vibrations on elastic rail support.

5. Summary

The determination of dynamic force for stress and strain calculations in the rail support elements caused by freight cars or locomotives with the axle base of less than 230 cm, especially for the track with modulus of rail support elasticity of up to 50 MPa, should take into account the additional pressure caused by the adjacent wheels. In other cases, the effect of the adjacent wheel is not essential. For modern passenger trains the level of the vertical force dynamic value does not depend on speed. The main factor for its determination should be the track vibrations as a result of its elastic properties and the presence of irregularities.

Acknowledgement

The author expresses gratitude to employees of “Railway Track Testing Laboratory” of the Dnepropetrovsk National University of Railway Transport, especially to Evgeniy Savluk and Olena Toropina for organizing and conducting the experiments.

References

[1] Danilenko EI, Rybkin VV: The computations rules of the railway track for strength and stability. TsP- 0117. Kyiv, Transport Ukrainy Publ., 64 p, 2004

[2] Chernyshov MA: Practical method of the track computation. Moscow, Transport Publ., 236 p, 1967

[3] Danilenko EI: Railway track. Structure, planning and calculations, interaction with rolling stock. Kyiv, Inpres Publ., Vol. 2. 456 p, 2010

[4] Fisher S, Eller B, Kada Z, Attila N: Railway construction Railway construction. Győr, 334 p, 2015

[5] Lichtberger B: Thack compendium. Eurailpress Tetzlaff-Hestra GmbH & Co. KG, Hamburg, 634 p, 2005

[6] Suiker AS, de Borst R: A numerical model for the cyclic deterioration of railway tracks. International journal for numerical methods in engineering, Vol. 57, pp. 441-470, 2003
DOI: 10.1002/nme.683

[7] Ma W, Chen T: Analysis of permanent deformations of railway embankments under repeated vehicle loadings in permafrost regions. Sciences in Cold and Arid Regions, Vol. 7, No. 6, pp. 645-653, 2015
DOI: 10.3724/SP.J.1226.2015.00645

[8] Remennikov AM, Kaewunruen S: Experimental load rating of aged railway concrete sleepers. Engineering Structures, Vol. 76, pp. 147-162, 2014
DOI: 10.1016/j.engstruct.2014.06.032

[9] Major Z: Longitudinal Behaviour of Embedded Rails. Acta Technica Jaurinensis, Vol. 8, No. 2, pp. 179-187, 2015
DOI: 10.14513/actatechjaur.v8.n2.367

[10] Bondarenko IO, Kurgan DM, Patlasov OM, Savluk VYe: Using of digital instrumentation equipment for experimental researches of track and rolling stock interaction. Science and Transport Progress. Bulletin of Dnipropetrovsk National University of Railway Transport, No. 37, pp. 124-128, 2011

[11] Connolly D, Giannopoulos A, Forde M: Numerical modelling of ground borne vibrations from high speed rail lines on embankments. Soil Dynamics and Earthquake Engineering, Vol. 46, pp. 13-19, 2013
DOI: 10.1016/j.soildyn.2012.12.003

[12] Krylov VV, Dawson AR, Heelis ME, Collop AC: Rail movement and ground waves caused by high-speed trains approaching track-soil critical velocities. Proc. of The Institution of Mechanical Engineers Part F-journal of Rail and Rapid Transit - PROC INST MECH ENG F-J RAIL R, Vol. 214, No. 2, pp. 107-116, 2000
DOI: 10.1243/0954409001531379

[13] Kouroussis G, Van Parys L, Conti C, Verlinden O: Using three-dimensional finite element analysis in time domain to model railway–induced ground vibrations. Advances in Engineering Software, Vol. 70, pp. 63-76 , 2014
DOI: 10.1016/j.advengsoft.2014.01.005

[14] Woldringh RF, New BM: Embankment design for high speed trains on soft soils. Proc. of the 12th Europ. Conf. on Soil Mechanics and Geotechnical Engineering (7.06-10.06.1999), Amsterdam, The Netherlands, Vol. 3, pp. 1703-1712, 1999

[15] Kurhan DM: Features of perception of loading elements of the railway track at high speeds of the movement. Science and Transport Progress. Bulletin of Dnipropetrovsk National University of Railway Transport, No. 56, pp. 136-145, 2015
DOI: 10.15802/stp2015/42172

[16] Koch E, Szepesházi R: Mélykeveréses technológia vasútépítési alkalmazásának lehetőségei. Sínek Világa, Vol. 2, pp. 9-14, 2013

[17] Petrenko V, Sviatko I: Simulation of subgrade embankment on weak base. Science and Transport Progress. Bulletin of Dnipropetrovsk National University of Railway Transport, No. 58, pp. 198-204, 2015
DOI: 10.15802/stp2015/49286

[18] Fischer S: Investigation of inner shear resistance of geogrids built under granular protection layers and railway ballast. Science and Transport Progress. Bulletin of Dnipropetrovsk National University of Railway Transport, No. 59, pp. 97-106, 2015
DOI: 10.15802/stp2015/53169

[19] Holder DE, Williams BA, Dersch MS, Edwards JR, Barkan CP: Quantification of Lateral Forces in Concrete Sleeper Fastening Systems Under Heavy Haul Freight Loads. Perth, Australia, 2015

[20] Manda KR, Dersch M, Kernes R, Edwards RJ, Lange DA: Vertical load path under static and dynamic loads in concrete crosstie and fastening systems. In 2014 Joint Rail Conference, pp. V001T01A025-V001T01A025, 2014
DOI: 10.1115/JRC2014-3832