Acta
Technica
Jaurinensis

Vol. 10, No.1, pp. 1-12, 2017
DOI: 10.14513/actatechjaur.v10.n1.405
Available online at acta.sze.hu

On Generalizations and Improvements to the Shannon-Fano Code

D. Várkonyi[1], P. Hudoba[2]

[1]Eötvös Loránd University, Faculty of Informatics, Department of Numerical Analysis
Pázmány Péter sétány 1/C, 1117 Budapest, Hungary
Phone: +36 1 372 2500
e-mail: varkonyidavid91@gmail.com

[2] Eötvös Loránd University, Faculty of Informatics, Department of Computer Algebra
Pázmány Péter sétány 1/C, 1117 Budapest, Hungary
Phone: +36 1 372 2500
e-mail: hudi1989@gmail.com

Abstract: This paper examines the possibility of generalizing the Shannon-Fano code for cases where the output alphabet has more then *2* (n) symbols. This type of generalization is well-known for the famous Huffman code. Furthermore, we will be looking at possible improvements to the algorithm, as well as other entropy based lossless data compression techniques, based on the same ideas as the Shannon-Fano code. All algorithms discussed in the paper were implemented by us in C++, and we will be illustrating our hypotheses with test results, made on an average performance PC.

Keywords: compression, lossless, entropy, Shannon-Fano

1. Introduction

Shannon-Fano coding [1][2] is an entropy based lossless data compression technique, which means that it is an algorithm for constructing a prefix code, in a way that a message coded in this will be shorter than normally. First, we need to define what we mean by “normally”.

Most coding schemes are developed explicitly for binary encoding, since modern computers use the binary system, however we will be encoding into an arbitrary system.

Let

$$l: \Omega \to L, |L| < +\infty \tag{1}$$

be a random variable, and the possible outcomes of l (the elements of the alphabet) letters. We will be working with finite alphabets, as shown above. Next we will be sampling this random variable n times, giving us a “message” of length n. This gives us:

$$l_1, \dots, l_n \colon \Omega \to L \tag{2}$$

This is a sequence of independent and identically distributed (IID) random variables. We will only be working with independent variables, however finding encoding schemes in cases where the letters are not independent of each other is of great interest, and it is a subject of dictionary based compression methods. These methods try to eliminate the redundancy of information in saying that the *i*th letter is *x*, when the *i*th letter is almost entirely (or entirely) determined by the previous (or following) letters. However this is not our case, we will be working with independent variables, meaning the *i*th letter is not at all determined by the other letters.

The concatenation of these letters will give us the message. So we arrived at:

$$m \colon \Omega \to M = L^n, m = l_1 \dots l_n \tag{3}$$

Now, we will be encoding our message in a “different alphabet” (which we will call symbol set) of smaller size then “originally”. Let our symbol set be:

$$|A| < |L| \tag{4}$$

The elements of the *A* will be called symbols. Since (4), we need to encode each letter with a sequence of symbols. The trivial solution would be to encode each letter in exactly

$$\lceil \log_{|A|} |L| \rceil \tag{5}$$

symbols. This way assigning symbol sequences to letters would be simple. Let

$$\tilde{l}_1, \dots, \tilde{l}_{|L|} \; and \; \tilde{a}_1, \dots, \tilde{a}_{|A|} \tag{6}$$

be an arbitrary ordering of sets *L* and *A* (all elements of the set occur exactly once in the sequence). Our “simple” assignment would then be:

$$f \colon L \to A^{\lceil \log_{|A|} |L| \rceil}, f(\tilde{l}_i) \coloneqq \tilde{a}_{\left\lceil \frac{i}{|L|/|A|} \right\rceil} \tilde{a}_{\left\lceil \frac{i - \left(\left\lceil \frac{i}{|L|/|A|} \right\rceil \frac{|L|}{|A|} \right)}{|L|/|A|} \right\rceil} \dots \tilde{a}_{\left\lceil \frac{i-(\dots)}{|L|/|A|} \right\rceil} \tag{7}$$

With this conversion between our two alphabets, our message encoded becomes:

$$m_e \colon \Omega \to M_e = A^{\lceil \log_{|A|} |L| \rceil n}, m_e = f(l_1) \dots f(l_n) \tag{8}$$

This is the trivial encoding, and thus by the “original length” of the message we will mean the number of symbols we can encode it in, using this technique, meaning the number of symbols needed to encode *1* letter, times the number of letters, so:

$$\lceil \log_{|A|} |L| \rceil n \tag{9}$$

Let us look at an example. Let our symbol set be binary, and let it consist of 0 and 1.

$$|A| = 2, \tilde{a}_1 = 0, \tilde{a}_2 = 1 \tag{10}$$

This is the case when we want to encode data on binary computer. We will encode a letter in as many symbols as we need, in order to be able to differentiate the letters (no two letters can have the same symbol sequence assigned to them). Naturally, we need

$$\lceil \log_2 |L| \rceil \tag{11}$$

symbols. Then we set the first symbol to 0 for each letter in the first half of the alphabet, and to 1 for the second half. We do this recursively on both halves, and we arrive at a binary encoding of our message.

This works great when the size of our alphabet is a power of $|A|$ (*2* in the binary case), and the letters have a uniform distribution, meaning that the *i*th letter in the message has the same probability being a certain element from the alphabet than any other. If either of these two conditions are not met, redundancy appears in the encoding.

If the size of the alphabet is not a power of $|A|$, symbol sequences will remain with no letters assigned to them. This is a “waste” of symbol sequences, a redundancy in the code.

If the letters’ distribution is not uniform, then *i*th letter is “partially determined” by its distribution alone, and thus stating it is *x* is “partially redundant”.

Both of these redundancies can be eliminated by clever assignment of symbol sequences to letters (*f).*

The only thing we need to keep in mind is that not only *f* has to be invertible (which is trivial), but the entire encoding scheme:

$$E: (\Omega \to M) \to (\Omega \to M_e), E(m) = m_e \tag{12}$$

Encoding concatenates the values of *f*, so if *f* is invertible but has in its range a symbol sequence which is a prefix (initial segment) of another symbol sequence, also in its range, the encoding can still be uninvertible. For example, let:

$$f(a) = 0, f(b) = 010, f(c) = 10 \tag{13}$$

With this choice, the encoded message *010* could have been encoded from *ac* or *b*.

To eliminate this possibility, and finding invertible encodings, we will only be constructing prefix codes (no symbol sequence in range of *f* can be a prefix of any other symbol sequence in range of *f*).

In most cases we will be given a message which we need to compress, which means that instead of knowing the actual distribution of *l*, we will have an empirical distribution.

2. The Shannon-Fano code

As stated previously the Shannon-Fano code is an algorithm for constructing prefix codes for data compression. It was developed for binary output (|*A*|=*2*). The algorithm follows 3 steps:

1. Order the set of possible letters into a list in descending order of their probability of occurrence (or number of occurrences). Place the list in the top node of a binary tree graph.
2. Cut the list into two sub-lists, in a way that the sum of the probabilities (or occurrences) in the two sub-lists are as close to each other as possible. Place

each sub-list in a child node of the list. Repeat this step recursively until no leaf node remains with more than *1* letters.

3. Assign bit sequences to the letters based on the binary tree, in the same way as Huffman's code [3]: for a given letter, walk down the tree to its node, getting a *0* from each left branch taken, and a *1* from each right branch.

The point of this algorithm is to maximize the entropy of each bit in the output one by one. Let us examine this statement a bit closer.

Let self-information of an event be the "surprise" this event means, the improbability of the event occurring:

$$I(l = 'x') = \log\left(\frac{1}{P(l='x')}\right) \tag{14}$$

Let the entropy of a random variable be the expected value of self-information the outcomes of the random variable could mean.

$$H(l) = E(I(l)) \tag{15}$$

This quantity will be maximal if the distribution of *l* is uniform.

Now, for a given symbol sequence which encodes a letter, the first bit determines whether the letter is on the left or the right branch of the root node. That first bit has maximum entropy if the distribution of its outcomes is uniform, or as close to uniform as possible. This is exactly what the Shannon-Fano code achieves by cutting the list in a way that the sum of probabilities (or occurrences) are as close to each other as possible.

After the first bit eliminated the maximum amount of uncertainty it possible could, the list cutting is called recursively, which maximizes the entropy of the second bit, and so on.

Altogether this maximizes the entropy of each bit, one by one.

Intuitively this algorithm could lead to an optimal code, but that is not the case. The problem is that while it achieves optimality bit by bit, it fails on a global scope.

If cutting a list at a certain place is optimal, it could be that after this cut we face much worse choices in the next step (in cutting the sub-lists) as if we had cut at another place, and overall we get a worse code then we could have, using a suboptimal cut at the first step. Basically the algorithm steers the code in the way that seems optimal in a local context, but fails to recognize the global optimum. In this way Shannon-Fano code is a greedy alternative to Huffman's code.

3. Generalization for *n>2*

Generalization of this type of compression to not only binary output alphabets could be of great interest, for example because since a lot of modern telecommunication methods use a *>2* symbol rate, which means a direct compression of the transmitted message could yield greater bandwidths.

Huffman's algorithm, while mostly discussed in the binary case, has an easy and well-known generalization for the *n>2* case. Huffman himself considered this case in his original paper in 1952. However, making this generalization for the Shannon-Fano code is not as straightforward.

In Huffman's algorithm, in binary case the algorithm needs to find the *2* lowest probability (or occurrence) node. For a general n, we simply find the n lowest probability (or occurrence) node, and the algorithm works the same as before.

Shannon-Fano code cuts a list into *2* pieces in a way that minimizes the difference of the sums of probabilities (or occurrences) between them. For a general n, we would need to cut the list into n pieces in such a way. This maximizes the entropy of the next symbol in the output, same as before. The problem that for $n>2$ there is no such thing as the "difference of the sums of probabilities (or occurrences) between them", there only exists differences of the sums of probabilities (or occurrences) between them.

We need to redefine what we mean by optimal cut. The underlying idea will remain the same, meaning that we want to set the distribution of the next symbol to as close to uniform as possible (and thus maximizing its entropy), for which we need a way to measure how "bad" a distribution is, how "far" it is from uniform.

Definition of this quantity is not trivial. We could use any number of such measures, however intuitively, two seems obvious:

- Let the maximum difference of the sums be minimal.
- Let the sum of the differences of the sums be minimal.

If we think of the differences between the sums as a collection of numbers (a vector) these measures correspond to the p-norms of $p = \infty$ and $p = 1$, the two extrema, thus it is likely that one of these methods will yield the best results.

We examined both of these methods, and tested them extensively. The results are not conclusive, but tend to favour the second method ($p = 1$). Before looking at the test results let us discuss the other problem in generalizing the Shannon-Fano code.

The code cuts a list into n pieces in each step. We need to think about what happens when a list has less than n letters. In the binary case, this could not occur, because the only way a node would have less than *2* letters, is if it has *1*, which is the stop condition of the recursive cutting, so it is not a problem. In the general case however this is very much a possibility. There are also two solutions to this problem that seems obvious:

- Let us leave empty nodes: If we have $m<n$ letters on a list, cut the letters into a separate sub-list each, and leave n-m lists empty. This means that symbol sequences would remain unassigned, which is clearly a "waste of symbol" in a way.
- Let us cut the list in a way that rules out that down the tree this problem could occur. To guarantee this we need to only cut the list in a way that creates sub-lists that have m letters on them, where

$$(m-1) \equiv 0 \; (mod\, n-1) \tag{16}$$

We will prove that this constraint guarantees that a full subtree can be created from the node in a constructive manner:

Clearly, if *m=1*, then a full subtree can be created, since the node itself will be the tree. *m=1* also satisfies the (16) equation. If we want to add another letter to the node however, we need to add not only one, but *n-1*, because adding only *1* would create empty nodes, which are filled in by adding a total of *n-1* letters. By induction we can conclude that a full tree can be created by adding a multiple of *n-1* letters, so (16) must be true.

With this method however we may not be able to choose the optimal cut, which results in suboptimal symbol distribution. The other problem that arises when choosing this method is that the first node in the tree might not be such, that a "full" tree can be created, meaning that there are inputs for which this method cannot be applied.

We also examined both two of these options, and tested extensively. The results are fairly conclusive, and favour the second method, where we avoid leaving empty nodes.

We tested our algorithms with random text generators, and books available on the internet as well. Table 1 shows a typical result we got, while testing on a 100 paragraph text generated by an online random text generator. The empty rows are cases where the algorithm can not be applied.

Table 1. Test results on a randomly generated 100 paragraph text

TEXT LENGTH	BINARY	3-ARY	4-ARY	5-ARY
ORIGINAL	364998	243332	182499	182499
HUFFMAN	260087	166335	132717	115893
S-F MAX	260559	172763	148761	132904
S-F SUM	260559	172763	148565	132904
S-F MAX S	260559	166335		119506
S-F SUM S	260559	166335		115893

The left column labels the different methods we tested. The original and Huffman rows show the original length of the text, and the length using the Huffman coding. The "S-F MAX" rows show the results we got while using the $p = \infty$ norm, and the "S-F SUM" rows show the $p = 1$ norm, and the rows labelled with an "S" at the and show the results while avoiding creating empty nodes, and the unlabelled rows show the results we got while creating empty nodes, when needed.

The test results we got (not only those shown above) clearly show that the "S-F sum S" method performs the best of this 4, almost reaching Huffman's code.

4. Non-ordering Shannon-Fano

We will now examine what happens when we skip the ordering step from the algorithm. Ordering the set into a list can keep us from finding the optimal cut. Thinking of a set of *m* numbers it is entirely possible, that the optimal cut (measured by either one of our methods) would create subsets that are not strictly increasing. For *n=2* examine the case when we have numbers $\{1, 5, 6, 10\}$. Clearly the optimal cut is $\{1, 10\}$ and $\{5, 6\}$, however we could not have reached this if we pre-ordered the numbers into a list, and only cut them accordingly, because this way we would have created lists $\{1,5\}$ and $\{6,10\}$. Figure 1 shows an illustration of the phenomenon, on the text "abbbbbccccccdddddddddd", and *n=2*.

The ordering step's main role is to make the cuts faster. Still thinking of the *n*=2 case, if a list is ordered, the optimal separation can be found by moving a separator from left to right between the letters, and stopping when the right side becomes smaller (in sum) then the left one. Not only so, but this advantage is "inherited" down the tree. The list cut in such fashion becomes two ordered sub-lists, for which finding the separation can also be done in this manner.

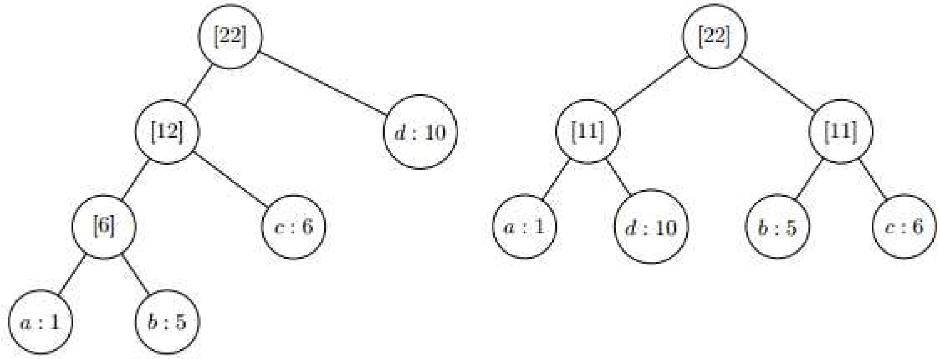


Figure 1. Ordering vs. Non-ordering algorithms

If we skip ordering, "finding the optimal cut" means that we need to cut a set into a disjoint union of *n* subsets, for which a certain measure of entropy is maximal.

Even in a simple case, for *n=2*, this is equivalent to the subset sum problem, which is NP-complete. In this case we are looking to find a subset of a set that sums up to as close to half the original set's total value as possible. Clearly, a simpler problem would be to find a subset of a set that sums up to exactly the original set's total value, which is a given number. This is the subset sum problem.

Overall, abandoning the ordering step will slow down the algorithm considerably, but could possibly improve the compression ratio.

In practice though, another factor comes into play. Even though ordering the set into a list can keep us from obtaining the locally optimal cut, it does not actually say anything about the globally optimal cut. It is possible that the non-ordered Shannon-Fano code, while getting better cuts locally, has a worse result globally. In fact, if we have a distribution which has a few very large numbers, and a lot of small ones (which means that the optimal compression ratio is high), the ordered Shannon-Fano keeps the small number of big numbers together, and creates a small subtree for them, and keeps a lot of small numbers together, creating a big, deep subtree for them. This guarantees the big numbers will be high on the tree, and the small ones will be on the bottom, getting us a good code. In the orderless version however, if the algorithm sticks one small number to a big one, that small number forces the big one down the tree another level, which is a great waste of information. So intuitively, it is entirely possible that the orderless algorithm gets worse results.

Figure 1 illustrates such a case. In this case the total output length using the ordered algorithm was 40, while using the non-ordered it was 44.

We implemented all our S-F variants in a non-ordering version, and tested them extensively. The tests conclude that not only the non-ordered algorithms run so slow, that it is practically impossible to apply them in any real-life situation, they result in a worse code (on a global scope) then their ordered counterparts for almost all inputs. These algorithms thus are clearly not worth investigating them further.

5. Distributing method

Up to this point we were constructing variants of the Shannon-Fano method by altering the way the algorithm chooses where to cut the set or list of the letters. We will now look at a completely different method of dividing a set of letters into *n* sets.

Let us create *n* empty sets, and assign our letters to one of these sets one-by-one. Take the letter with the highest probability (or most occurrences) and assign it to the set that has the least probability (or occurrence) in sum. Repeat this step until all of our letters have been assigned to a set. We are thus distributing the letters between the sets, in a way that keeps the distribution fairly uniform.

This method however has a defect. In a given step, if all but one set (the ith) have been assigned $1, n, n^2, \dots$ letters (the sets have a power of n number of letters), then it is almost surely better to assign the next letter to the ith set, regardless of the values on the nodes. The reasoning behind this, is that on the next level down the tree, the “full” nodes will leave no empty nodes after their division. A node with 1 letter becomes a leaf, a node with n letters becomes a 2 level full tree, a node with n^2 letters becomes a 3 level tree, and so on. Adding any letter to a full node would trigger building another level somewhere in their subtree, which is clearly a waste, because it lowers other letters in the full tree, giving them a longer symbol sequence. If we assign the next letter to the ith set however, no other letters have to be forced down a level, and we get a better code.

There are exceptions to this argument. If the empty node is so far down the tree, that assigning the next letter there would result in such a long symbol sequence, that it is better to lower some letters higher up in the tree, and assign the next letter to the empty node we created, then this argument would be invalid.

This however is a very unlikely scenario, and happens only in extreme cases. In most cases, it is worth it to always fill the sets to “full” first, and then choose one to spoil. It is also possible to correct this flaw in the algorithm, by only assigning a letter to a non-full node, if down the tree it will not give us worse results overall then we would get if we spoil another full node. This however would require that we know information about the full tree in advance. It is possible to “try out” the assignment to the non-empty set, and calculate in advance if it would be a better choice, but that would be less of a Shannon-Fano code, as it would use information on the full tree to make a decision, and not only information on the level in question. It would no longer be an algorithm which builds a tree from top to bottom, by splitting its branches recursively, which is in essence what the Shannon-Fano code does.

Figure 2 illustrates a case where dividing into sets that have a power of *n* letters clearly shows its advantage. Our text is "aaaaaabbbbbccccdddee", and *n=2*. On the left side, the optimal division of the first list caused empty nodes to be created on the second one, while on the right side no empty node is created. Using the first method we would need 34 symbols to encode our text, while using the second one this number is only 29.

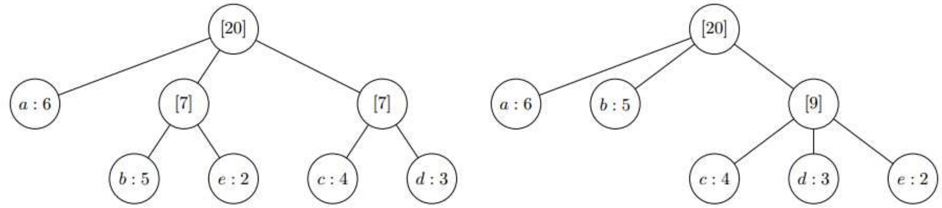


Figure 2. Distributing methods

We implemented both versions of the distributing method, an algorithm where we do not care about empty nodes (labelled "S-F DISTRIBUTING"), and one where we fill to "full" first (labelled "S-F DISTRIBUTING S"). Both algorithms have been tested extensively, and overall we found that the latter one performs better in almost all cases. Table 2 shows our test results, on the same input text as the previous tables.

Table 2. Distributing method

TEXT LENGTH	BINARY	3-ARY	4-ARY	5-ARY
ORIGINAL	364998	243332	182499	182499
HUFFMAN	260087	166335	132717	115893
S-F DISTRIBUTING	268187	179727	141954	125097
S-F DISTRIBUTING S	268187	171117	132817	117311

6. Pushdown method

As stated previously, the essence of the Shannon-Fano coding technique is to build up a tree by starting from a root node, and recursively splitting it into branches. For our last method we will be stepping outside this framework somewhat.

The distributing method assigns each letter to a node in the next level one by one, and after it is finished, it is repeated on all nodes created in the next level, recursively. Let us construct a similar method, one which assigns letters one by one, but assigns them to a node anywhere in the tree, not only the next level. This way no recursive calling will be required, as we place the letter in the tree, not only decide on which branch will it reside.

This is no longer a Shannon-Fano method in the way Shannon-Fano methods work level by level. It is however still a Shannon-Fano method in the way that it will make locally optimal choices.

The only thing we need to work out is how we assign a given letter (*l*) to a node in the tree. We can:

- Place it on an empty node on the *i*th level
- Divide a node on level *j* in the tree which has letter *L* assigned to it into another level of *n* nodes, assign *L* to one, *l* to another, and thus create *n-2* empty nodes.

Let us consider how many extra symbols would appear in our output if we would do these steps:

- If we place it on an empty node on the *i*th level, letter *l* would be assigned a symbol sequence of length *i*. If letter *l* has *P(l)* occurrences that would mean *i*P(l)* extra symbols.
- If we divide an existing node, that means that *l* will be on the *(j+1)*th level, which gets is *(j+1)*P(l)* extra symbols, plus *L* has been lowered a level so we need *1* extra symbol for each occurrence of *L*, resulting in a total of *(j+1)*P(l)+P(L)* extra symbols.

Obviously we should place the letter where it causes the least extra symbols in our output. Thus we have to calculate these values for all possible places, and choose the best option.

This is still not the best algorithm we could construct this way, because it does not take into account that dividing an existing node creates many empty nodes, which are potentially better places for future letters. Constructing an appropriate "weight" however for this effect is a difficult task, so we will not get into that. Instead, we will try to model this in another way.

Let us assign *n-1* letters at a time. In each step, we will gather *n-1* letters from the unassigned ones, choose where to put them, by calculating the weight of each possible place in a similar weight, and assign them there. This means that we do not need to weight empty nodes, because we will only add *n-1* to a "full" tree, which will remain a "full" tree. At the end, there may be extra letters left, if the total number of letters is not *n+i(n-1)* for some *i*. In this case, we will assign the rest of the letters one by one, as before.

This is basically the same algorithm as if we would have weighted all empty nodes at 0. If we would have done so, then after breaking a leaf node into a sub-tree, the next *n-2* letters (a total of *n-1*) would have been assigned to the empty nodes created. So we would assign letters not one by one, but *n-1* by *n-1*.

We implemented both versions of this method, and tested them extensively. Table 3 shows the test results we got on the same input text as we used before.

Table 3. Pushdown method

TEXT LENGTH	BINARY	3-ARY	4-ARY	5-ARY
ORIGINAL	364998	243332	182499	182499
HUFFMAN	260087	166335	132717	115893
PUSHDOWN	269426	177533	141245	124005
PUSHDOWN N-1	269426	168401	133356	115893

7. Conclusion

In this paper we began by introducing the problem of entropy based encoding in a general case, not only binary output. We introduced the Shannon-Fano coding for binary output. Next, we examined the possibility of generalizing the Shannon-Fano code for arbitrary output alphabet, in a similar manner, as one would generalize Huffman's code. We saw that there are some difficulties in doing so, for which we gave several solutions ("S-F MAX", "S-F MAX S", "S-F SUM", "S-F SUM S"). Afterwards we looked at possible improvements or alterations to the Shannon-Fano code. First we looked at non-ordering versions of the previously constructed algorithms, which yielded worse results than the ordered versions. Then we looked at alternatives to these algorithms, namely the distributing and pushdown methods ("S-F DISTRIBUTING", "S-F DISTRIBUTING S", "PUSHDOWN", PUSHDOWN N-1"). These algorithms performed well in our tests, but the best algorithm still seems to be "S-F SUM S".

Based on this judgement we ran a longer test, on a bigger input, with bigger alphabet, this time only with our "winner algorithm". Our randomly generated 100 paragraph text, which was used during testing for all the algorithms had 41 distinct letters, and had a size of 60 833 bytes, while the final test, ran only on the "S-F SUM S" method had 256 distinct letters, and had a size of 416 256 bytes. This test file was the executable file of the program itself which we used to test all our algorithm. Table 4 shows our test results.

Table 4. Final result

TEXT LENGTH	BINARY	3-ARY	4-ARY	5-ARY
ORIGINAL	3330048	2497536	1665024	1665024
HUFFMAN	2202062	1402926	1117106	963429
S-F SUM S	2207253		1136380	

Overall, we can conclude that this method is the most efficient generalization of the Shannon-Fano code for arbitrary output alphabets, and performs very closely to Huffman's code. The method however does have the disadvantage that it does not work on all inputs, and it runs slowly, because we used a brute force approach for finding the optimal cut. Subject of further research into the area could be the construction of a non-brute force algorithm for finding this optimum.

References

[1] Shannon CE: A mathematical theory of communication, Bell System Technical Journal, Vol. 27, pp 379-423, 1948

[2] Fano RM: The transmission of information, Technical report no. 65, Research Laboratory of Electronics, Massachusetts Institute of Technology, 1949

[3] Huffman D: A method for the construction of minimum-redundancy codes, Proceedings of the IRE, Vol. 40, No. 9, pp 1098-1101, 1952

[4] Cover TM, Thomas JA: Elements of information theory, 2nd edition, New Jersey, John Wiley & Sons, Inc., 2006, ISBN: 9780471241959

[5] Jones GA, Jones JM: Information and coding theory, London, Springer, 2000, ISBN: 9781852336226

[6] Yeung RW: Information theory and network coding, London, Springer, 2008, ISBN: 9780387792330

[7] Sayood K: Introduction to data compression, 2nd edition, London, Academic Press 2000, ISBN: 9781558605589

[8] Sayood K: Lossless compression handbook, London, Academic Press 2003, ISBN: 9780126208610

[9] Witten IH, Moffat A, Bell TC: Managing gigabytes: compressing and indexing documents, and images, 2nd edition, London, Academic Press, 1999, ISBN: 9781558605701

[10] Salomon D: Data compression: The complete reference, London, Springer, 1997, ISBN: 9780387982809

[11] Salomon D, Motta G: Handbook of data compression 5th edition, London, Springer, 2010, ISBN: 9781848829022

[12] Drozdek G: Elements of data compression, Pacific Grove, Brooks/Cole publishing, ISBN: 9780534384487

Acta
Technica
Jaurinensis

Vol. 10, No. 1, pp. 13-20, 2017
DOI: 10.14513/actatechjaur.v10.n1.410
Available online at acta.sze.hu

System Approach for Strategic Planning in Transport

B. Horváth, B. Gaál

Széchenyi István University, Department of Transport
Egyetem tér 1., 9026 Győr, Hungary
Phone: +36 96 503 494
e-mail: gaalb@sze.hu

Abstract: The paper gives an overview on the current status of the connection between city and transport. After a short theoretical introduction on the effect of different city development strategies it shows the importance of transport in the life of a settlement. Through international best practice, there are some good examples on the holistic planning of a city.

Keywords: transport system planning, urban planning, mobility management

1. Introduction

István Széchenyi is esteemed as the founder of the Hungarian Academy of Sciences or the builder of Chain Bridge. However, in the field of transportation his biggest contribution was the creation of the first transport policy. [1] One of the most important and relevant statement of the policy was the following: *"Transportation is not constituted as the core of the state, but it has the same effect as the veins in a living body"* and *"…there is no such a sector in our public and private economy, which is not affected [by transportation]…"* [2]. Nowadays, in Hungary, these two statements have been forgotten or at least overshadowed by other factors. Sadly, it negatively affects the livable, sustainable city life.

This article, through some ideas and examples, would like to draw attention to the inseparability of the city and its transportation system. During the planning of cities we (should) consider the effects and demands of the developments.

In this article we use the word transportation in its broadest meaning it includes transportation of people and goods, public and individual, and motorized and non-motorized transportation. System approach also means we do not emphasize certain solutions rather we seek equilibrium in the multitude of solutions. The importance of the topic is indicated by the 7th consecutive victory of Vienna in the quality of living rankings in 2016 which victory was partially achieved thanks to its transport system.

2. City and transportation

Széchenyi had discovered already in the 19th century the transport economics principle that transport demand is a derived demand, that is, it is generated by the spatial separation of socioeconomic processes. Subsequently, the problems of transportation systems cannot be handled by only the tools of transport planning or traffic management. Csaba Orosz [3] had similar concept when he said that the best (cheapest) trip is which does not happen. Hörcher [3] approached this theory from economics point of view.

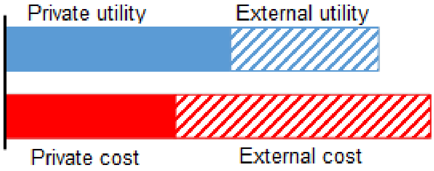


Figure 1.: The real and putative utilities and costs of trips [4]

He expressively defined the economic axiom that a trip results in negative utility for passengers (Fig. 1.). If the cost of a journey is less than the utility of the reached activity the journey will be realized. For example, when one is looking for a job he will ponder on the distance from where commuting is still worth. If the transportation system, regulatory environment and the structure of the city is such that one has to travel a lot, but he can do it cheaper (as cheaply we meant the complex of price, time and other factors) then one will do the journey. However, this journey costs not only him, but also society as it can be seen in figure 1. This means when an employee regularly takes long rides to the workplace not only he inflicts cost to himself but also for society. If we encumber travelling, for example by charging the traveller for fully or partially the real cost of travel then this could force the change of travel behaviour (for example, they move near to their workplace or switch to public transport – if there is any). This simple example also reveals that the first step in properly managing the transport system is the examination of transport demand (why people travel?) and the planning or development of the transport system as a technological system only could follow this phase.

Transport demand depends on different places has to be reached by the user and their spatial distribution as most transport models consider it in their trip generation phase. Traditionally simple linear correspondence can be demonstrated between the number of inhabitants, schoolchildren, jobs, shops of an area and departing/arriving trips.

3. Transportation effects of districts and facilities

Nowadays, in the age of mobility demand for transport rising like never before. In the light of this fact, it is worth considering the development of districts or facilities. When we speak about the development of districts we have to separate the development of neighbourhoods and economic zones. In our days, economic zones are mainly implemented as mono-functional greenfield or brownfield investments. The primary aspect of the location choice is the transportation needs of raw materials and/or (semi)finished products. What about the employees? How can they reach their workplace? Is this question emerged during the designation of development areas? Why is it important? There are two answers: time and money. While the local authorities battle for investors by designating development areas at the same time it is a serious problem for them to finance the public transport lines of these areas. After all, only few, big employer could allow itself the organized transport of employees. Is there anyone who considers during the development process the financing of future public transport's operation?

It is may be demagogic to say that move dwellings closer to industrial areas as no one would like to live in the vicinity of one. However, it is highly important to take into account this – in the time of the planning process – latent travel demand.

Figure 2.: Typical subdivision (Source: maps.google.hu)

Other similar problems occur with subdivisions. In this case the main question is the evaluation of future traffic between the development area and other parts of the city. Did anyone examine the trip generation effect of the new quarters in the master development plan? Did it occur to anybody that not only public utilities, but road network also have to be able to support the new population and its activity? As we already mentioned it is the age of mobility and these subdivisions function like classic commuter towns. Furthermore, because of the low intensity of built-up it generally lacks efficient public transport system.

As a result, a 10 apartment complex with an average of 30 residents (given the average 1.5 people per car) means 20 cars. With ten such a building it is 200 new cars participating in traffic. At night these cars need parking place also. In Hungary OTÉK (Decree No. 253 of 1997 (XII. 20.) of the Government on national urban development and building requirements) determines the minimal parking place per households, which is exactly one. Some local government raises the minimal parking place to 1.5 per household, but in our little example, there is another five which remains on the road (assuming that everyone bought the parking lot marked out for him and uses it). Residential parking lots are under-proportioned and as they have been sold separately from apartments (and many do not buy them) they tend to be empty and cars of the residents are stored on the road surface.

Although this could lead to the higher usage of public transport, but the system – because of the abovementioned inefficiency – mostly had not been established in due form to compete with private transport.

Another kind of problem occurs when the intensity of built-up is high, but during the planning process public transportation was not taking into consideration (lack of stops and turns, non-passable roads).

It is important to mention other high traffic attraction facilities like malls or stadiums. They have a dual demand as transport have to serve both the facilities (shops) and the visitors and not only with parking lots.

To summarize this train of thoughts it is important to state that in order to a development fully functions its accessibility is also decisive, but in Hungary only parking is regulated by OTÉK [5].

4. Transport demand and spatial structure

Modern society based on mobility and transport possibilities. Transport demand originates from the spatial division of economic activities (spatial division of labour). Earlier in history every function (housing, workplace, service) were present in a very limited space. In modern times they are divided by larger and larger distances. This is originates from the era of the industrial revolution. At that time earlier manufactures gave up their place to big factories. These factories had a considerable environmental impact (noise, air pollution) thus dwellings built up not next but near to them. This structure projected the posterior homogeneous districts, but due to low distances traffic problems were small-scaled.

By evolution of technology, especially transport technology, people could take longer and longer journeys, regularly and dependably. Thus, concentrated industrial areas and living quarters situated increasingly further away from each other generating more and more transport needs. At the same time, the concentration of workplaces and dwellings resulted in a relatively homogeneous travel structure which could be satisfied by public transport services. Besides residential areas contained the necessary place of services (day-nursery, kindergarten, school, pharmacy, market, stores, etc.) thus in everyday life usage of the transport system other than the commute to work was not necessary. Contrarily, at present subdivisions prevail which on one hand do not offer these primary functions, on the other hand, they do not get sufficient public transport service. Reasons could be the lack of critical mass or hiatus of plans concerning transportation needs. Because all of this transport acts the part of infrastructure, meaning we build roads in order to let population access services they cannot get on the spot. The question is which solution is more efficient in the long run.

5. Planning examples

Thereafter, we would like to present some good practice examples in the field of transport-aware urban planning.

Transit oriented development (TOD) is a desirable possibility for rapidly growing urban areas. The conception is simple: let’s make a self-sufficient town or district with a relatively dense built-up of housing, services and workplaces around high capacity stations of transport lines (mostly train or light train). Combined with a pedestrian friendly environment this kind of development could lessen private car usage and motorization.

Bogota, Columbia, started its BRT (Bus Rapid System) project at the end of the nineties. Its precursor was Curitiba, Brazil. Opposite to the Brazilian practice the management of Bogota concentrated on the development not only the city centre and the BRT system. Of course they implemented the separate bus lanes on main roads with attractive stops. At the same time they also implemented a pedestrian and cycling infrastructure development connected to the system even in the poor districts. This made the new system easily accessible even by foot. After the inauguration, the system was a huge success. The number of accidents was lowered, pollution decreased, one-tenth of the users chose the Transmilenio (this is the name of the system) over their cars, even criminal statistics changed for the better. Despite this success Transmilenio have not had such a land use altering effect like the one in Curitiba, albeit along the lines services met more favourable terms and opportunities and real estate prices rose. Of course, many factors contributed to the success of the change of the transport system. The Transmilenio system was well-planned (P+R at peripheral stops, high accessibility of stops by foot and bike, joint infrastructure) but the cycling infrastructure had a more astounding impact. More than 200 km dedicated cycling path was built and the share of cycling rose from 0,9% over 5%. Popularity of cycling was influenced by favourable conditions like relatively flat surface (Bogota is situated in a valley in the Andes), mild weather all year round and mixed land use which caused the under 10 km average travel length. Success was strengthened by further actions like car-free days, road closings on Sundays, car usage restrictions, etc.

The mature form of TOD is present in Europe, its cradle. Development plans of Copenhagen (“Finger Plan”) or Stockholm (“Planetary Cluster Plan”) are early representatives.

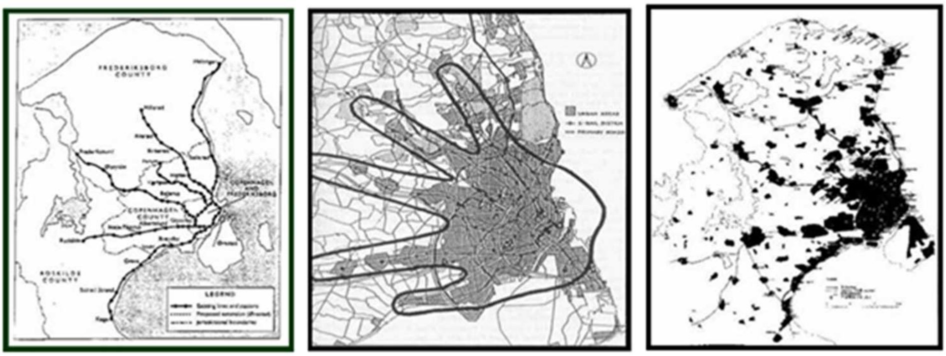


Figure 3.: The realization of the “Finger Plan” in case of Copenhagen [6]

Such a suburban development of Stockholm became an international model [7] primarily as a green city[8]. Hammarby Sjöstad was an abandoned industrial area. After Sweden lost in the tender for 2004 summer Olympics it was developed as a brown-field investment. Hammarby is situated only 3 km from the downtown of Stockholm. It offers densely built-up, but attractive environment (nearly 20.000 inhabitants). During the planning process many viewpoints were taken into consideration like mixed land use, high level accessibility and attractive streets for pedestrians and cyclists. Most services were placed along the main transport axis, other shops and restaurants were scattered in the residential areas. On the whole 30 percent of the suburb is industrial and business area. Despite environmental friendliness car infrastructure is also well-developed. There is 0.7 parking places per flat and the southern motorway is next to the suburb. To counteract these effects the city has a good public transport connection with Stockholm through bus lines, high capacity tram and ferries. Above all, there is a car-sharing system providing cars for members. Thanks to all of these in 2010 the modal-split was the following: 52% of the trips were made by public transport, 27% on foot or by bike and only 21% by car.

Freiburg had a similar model project [9] [10]. Vauban (5.000 inhabitants), which is also 3 km away from the city centre, was built as a pilot program for modern, energy-efficient, green cities. The district was built in such a fashion that the workplaces and services could be reached on foot or by bike. It was also the main consideration during shaping the roads. Car traffic was encumbered by hamstring the transit and speed-limits (30 km/h on the main road and 5 km/h on other roads). Bus service reaches only the edges of the district, while the tram line connecting the district to the downtown is in the main axis. Further aggravating for car owners is the strict parking regulation. Although the district is not outright car-free zone (or according to [11] it could be considered as a type of car-free development) and residents can own automobiles, at the same time they must park them in communal parking lots situated on the edges of the district. Residents have to buy their parking space and also pay further regular fee for it. While results are considered good (16% of the trips made by car, 19% by public transport, 64% on foot or by bike – which is better than the Freiburg average) yet problems arise in the run. Most people support the idea of car-free, liveable environment as long as they do not have to abandon theirs. Many inhabitants tried to avoid the high fees by disavowing or un-admitting their cars parking those in the surrounding districts aggravating their situation.

The project of Vauban especially well-documented from a transportation point of view, see [12][13][14][15].

Media City, Manchester, was a phase of one of the largest urban rehabilitation project in England. In place of old docklands new, attractive business and cultural centre has been built to revitalize the run-down part of the city. Media City is the example of joint development of urban and transport concepts. Early in the preparation stage it was clear that recent links will not cope with the anticipated extend of traffic, thus, in parallel, their development is much needed. Investors and the regional development agency gave 25 million pounds to local transportation association to improve transport connections. As a result a new spur from the existing tram line was built and new trams were purchased. It made possible the frequent service to the centre of Manchester. On the other hand tram line also ensured the accessibility of residential areas. Provident development was especially important for the local transportation association as their business is fully commercialized, they do not get subvention and have to finance and invest from their own profit.

However, new tram line alone was not sufficient. The neighbouring district, Salford, was approachable only by car, thus the parties concerned agreed upon starting a new bus line. As there is no subvention in the UK for these services, but it has been deemed necessary the local government and the university agreed to sponsor the line for 5 years (when it is expected to be self-supporting). Beside public transport non-motorized forms was also focused on. Attractive pedestrian and cyclist infrastructure was important contributors to the success (pedestrian bridges over the docks, B+R facilities). Investors were motivated to build them by the regional development council as they give their consent to continue the development only if investors possess such a plan which guarantee that 45% of the trips are made by other modes than a car.

Other local examples are intermodal centres. These are mainly terminals, places with excellent connection. They are frequently situated in run-down downtown areas where they can be the centre of renewal. During the renewing process these places previously with solely interchange function became multifunctional communal areas. In Hungary, similar albeit unsuccessful attempts were KÖKI terminal and West End. The fail of these initiatives were because while situated next to train terminals, they were not an organic whole and their function did not strengthen each other.

6. Conclusions

As a conclusion it can be stated that transportation should have to make a greater role in regional and urban development. Planners should be aware the fact, that transportation is a derived phenomenon which is shaped by the development. If we mix the functions within an area, or link the homogeneous areas by adequate links, less and easier to handle transport demand will appear. This means more liveable environment, clear air and less social externalises.

Our settlements could be made sustainable and desirable in the long run only by the triad of good regulatory environment, transport-oriented regional development and transport development aiming for healthy equilibrium of modal split.

References

[1] The 1848 XXX. On the tasks of the Minister for the vehicles (in Hungarian)

[2] Széchenyi I: Proposal for settlement of the Hungarian transport case, 1848 (in Hungarian)

[3] Orosz Cs: The possibilities of influencing travel mode choice in Budapest, Városi közlekedés 1993/2. pp. 88-97 (in Hungarian)

[4] Hörcher D: The optimal congestion: pricing and capacity management in public transport, in Közlekedéstudományi Konferencia Győr 2016. Universitas Kft., pp. 66-87 (in Hungarian)

[5] 253/1997. (XII. 20.) Government Decree, national town planning and building requirements (in Hunagrian)

[6] Cervero R: Public Transport and Sustainable Urbanism: Global Lessons, 2006. UC Berkeley: University of California Transportation Center http://escholarship.org/uc/item/4fp6x44f (2016.07.01.)

[7] Taylor I, Sloman L: Thriving Cities: Integrated Land Use and Transport Planning, 2011.

http://www.urbantransportgroup.org/system/files/20112706ptegThrivingCitiesReportforWebFINAL.pdf (2016.07.01.)

[8] Brogren M, Green A: Hammarby Sjöstad–an interdisciplinary case study of the integration of photovoltaics in a new ecologically sustainable residential area in Stockholm, 2003. Solar Energy Materials & Solar Cells 75, pp. 761–765 http://dx.doi.org/10.1016/S0927-0248(02)00133-2

[9] Williams J: Can low carbon city experiments transform the development regime? 2016. Futures 77, pp. 80-96. http://dx.doi.org/10.1016/j.futures.2016.02.003

[10] Melia S: On the Road to Sustainability: Transport and Car-free Living in Freiburg 2006. University of the West of England Faculty of the Built Environment www.stevemelia.co.uk/freiburg.doc (2016.07.01.)

[11] Melia, S, Parkhurst G, Barton H: (2011) Carfree, low-car - what's the difference. 2011. World Transport Policy & Practice, 16/2. pp. 24-28. ISSN:1352-7614

[12] FitzRoy F, Smith I: Public transport demand in Freiburg: why did patronage double in a decade? 1998 Transport Policy 5: pp. 163– 173 http://dx.doi.org/10.1016/S0967-070X(98)00024-9

[13] Nobis C: Evaluation des Verkehrskonzeptes im autoreduzierten Stadtteil Freiburg-Vauban 2003. Informationsnetzwerk „Wohnen plus Mobilität“ - Fachbeitrag Nr. 33 www.mobilitaetsmanagement.nrw.de/cms1/download/fb_33.pdf (2016.07.01.)

[14] Nobis C: The Impact of Car-free Housing Districts on Mobility Behaviour — Case Study. in Beriatos E., Brebbia C.A, Coccossis H, Kungolos A (eds): Conference on Sustainable Planning and Development 2003.WIT, Dorset, pp. 701–720. ISBN: 978-1-85312-985-8

[15] Broaddus A: A Tale of Two Eco-Suburbs in Freiburg, Germany 2010. Journal of the Transportation Research Board 2187 pp. 114-122 http://dx.doi.org/10.3141/2187-15

Vol. 10, No.1Y, pp. 21-34, 2017
DOI: 10.14513/actatechjaur.v10.n1.411
Available online at acta.sze.hu

NeuroCar Virtual Driving Environment: Simultaneous Evaluation of Driving Skills and Spatial Perceptual-attentional Capacity

H. Hämäläinen[1], F. Rashid Izullah[1], A. Aho[1], M. Koivisto[1], T. Laine[1], P. Qvist[2], A. Peltola[2], P. Pitkäkangas[2], M. Luimula[2]

**[1]University of Turku, Centre for Cognitive Neuroscience,
Turku Brain and Mind Center
Assistentinkatu 7, 20014 University of Turku, Finland
Phone: +358 400711364
e-mail: hhamalai@utu.fi**

**[2]Turku University of Applied Sciences, Turku Game Lab
Joukahaisenkatu 3, 20520 Turku, Finland**

Abstract: We describe here a simple, inexpensive and effective system for simultaneous evaluation of a subject's driving ability and spatial auditory and visual perception and attention. It consists of a commercial steering wheel and virtual glasses and a program for driving on a two-lane road with curvatures at about 100 km/h speed, and simultaneously reacting by pressing two buttons attached to the steering wheel to randomly delivered uni- and bilateral auditory signals via earphones and light dots appearing in the peripheral visual field. Three different difficulty levels of the task were applied in randomized counterbalanced order, each session of 2 min duration. The results of 25 young (17-45 years) and 20 elderly (47-96 years) healthy participants demonstrate the tendency for simultaneous right side spatial perceptual/attentional bias and the left side driving bias especially in the elderly participants.

Keywords: driving ability, spatial perception, attention, bias, age

1. Introduction

We have shown in our previous studies that humans are hardwired to a rightward bias in spatial perception and attention - this concerns at least most right-handers [1- 3]. This bias is strongest in children and elderly, but occurs also in young when under cognitive stress. We have presented a hypothesis and proposed a model, based on latest findings in cognitive neuroscience, according to which this age- and cognitive load-dependent spatial bias is due to the lateralization of attention mechanisms in the human brain, and slow development in the young and later decline in the old by the executive functions which are responsible for our skills for cognitive control, e.g. (in)voluntary attention, working memory, and purposeful, goal-directed behavior. The changes of the spatial bias as a function of the lifespan - distinct in childhood and old age - well reflects the interplay between the attentional-cognitive control mechanisms [4].

Since we only studied this bias in laboratory the question remains, whether this bias can also be demonstrated in more ecologically valid conditions, i.e. traffic. Therefore we constructed a simple system with steering wheel with inserted right and left buttons for responses. Virtual 3-D environment for the driving task was created with virtual glasses (Oculus Rift) and visual stimuli were applied to the periphery of the visual field and auditory stimuli were delivered via earphones. With this system it is possible to mimic and even replicate the previous laboratory experiments [2, 3] simultaneously while the subject is driving the imaginary car in along the road and in different driving conditions. This simple system allows us to determine the subject's capacity to 1) drive (keep the vehicle on the road/lane), respond to the auditory and visual spatial stimuli while driving, and the mutual effects of these two tasks on each other at different difficulty levels.

With this system we aim at an international database-based evaluation platform, where one can very quickly screen whether the subject in case drives and perceives the spatial stimuli properly according to the age, handedness, gender, etc. specifications. If not, the system warns the experimenter, and the subject needs more specific examination.

Based on thorough piloting we ended up in running three different difficulty levels to participants from all age levels. This is due to the fact that the variability in spatial perception, speed, and (quite unexpectedly) the speed and familiarity in button press technique is today so large, that if having only one difficulty level we would have ended with numerous younger participants with a ceiling effect and numerous older participants with interruption of the task because of impossibility of accomplishing the task as the other extreme.

In this report we will describe the properties and application of the system and the first results on effects of age on driving performance and simultaneously measured spatial attention perception and attention. Because we also wanted to check the occurrence of the right bias in our elderly participants, only right-handed participants were included in the present population.

2. Methods

2.1. Apparatus and stimuli

NeuroCar system

NeuroCar system was first introduced in (Luimula et al., 2015). Our multidisciplinary team has developed the first prototype of the virtual evaluation tool to be used for driving acuity and spatial perceptual capacity test in a virtual environment. This prototype consists of Oculus Rift Development Kit 2 data glasses, off-the-shelf-headsets, Fanatec Porsche 911 GT3 RS V2 Wheel US attached with Arduino response buttons, and software developed with Unity 5 game development platform. Oculus Rift data glasses has been used for 3-D presentation of the road and spatial peripheral visual stimuli, and headsets for presenting spatial auditory stimuli (the data glasses can be replaced for certain purposes by video screen).

The developed prototype has a user interface designed for the test subject but it offers for the researcher also administrative features. The difficulty level of the driving and spatial perception can be controlled and adjusted to the demands of the performance level of the driver (can be compared against normative data in a large international database later, backend system explained later in this chapter). Parameters to be varied are at the moment a speed of the vehicle, a rate of the spatial stimulus application, a visibility of the road, difficulty of the road (curvature, and width), and intensity of the stimuli (simultaneous evaluation of functioning of the visual and auditory systems). At the moment, administrative panel contains also features that will help to establish right parameters for the final product to simulate realistic driving experience.

Driving acuity can be determined during the driving task itself. Our virtual evaluation tool measures the rate and amount of driving errors (i.e. driving out of lane, direction of errors, and time spent outside the lane), where it is possible to assess the level of driving performance and the possible bias to the left or right. Also the relation of the driving errors to the spatial visual and auditory stimuli will be determined (the induction of the errors by the spatial stimuli, e.g. whether the response to right stimuli induces driving error to the right or left).

Spatial perceptual capacity can be determined during the driving with controlled series of visual and auditory peripheral spatial stimuli in separate trials or mixed together. With the very same stimulus series it has been shown distinct deterioration and biasing of the perceptual abilities in the healthy elderly population in laboratory conditions without any background driving task. The participant's task is to respond as quickly as possible to stimuli by pressing the left or right button (attached to the steering wheel) or both, to stimuli delivered from the left or right direction or from both simultaneously, correspondingly. The accuracy of responses and reaction times will be determined. This the most objective way of determining the degree of perceptual biases in the driver's perceptual capacity.

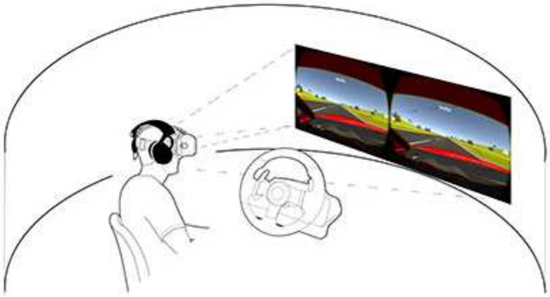


Figure 1. System description of the NeuroCar system in this study.

NeuroCar version reported in this paper is designed only for driving inspections so far. The training environment is at the moment under development. In this study, the driver is steering a virtual car which can be steered with a haptic steering wheel. The test subject (the driver) has headsets which will simulate 3D auditory environment. In addition, the test subject has Oculus Rift data glasses which will show for the driver virtual environment including road, simple terrain, and visual signals. While driving the test subject has to react on visual and auditory disturbances by pressing buttons attached to the steering wheel.

In the current prototype (Figure 1), driving environment is visualized in Oculus Rift data glasses which enables an immersive driving feeling. The camera view is showing the driving view from the car inside (Figure 2). The driver is able to see 3D virtual environment. Data glasses are reacting on the driver's head movements to increase the immersion. Visual disturbances are visualized peripherally 50 milliseconds and the driver has to react on those by pressing buttons attached to the wheel (defined in the system as mouse buttons). Similarly the driver has to react on auditory disturbances by pressing the buttons while hearing 3D sound stimuli from the headsets.

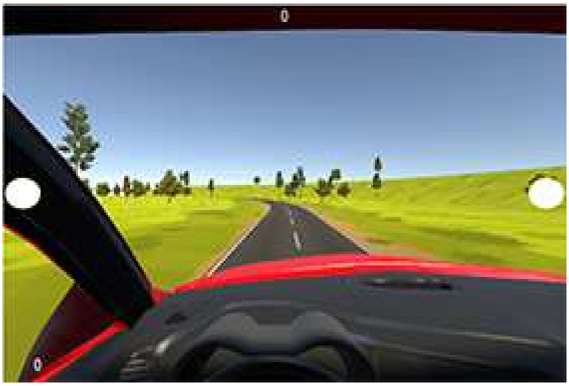


Figure 2. A screen capture of drivers's perspective in Oculus view.

The buttons has been implemented attaching external buttons to the steering wheel controller. By pressing these buttons left or right mouse buttons clicks will be activating. These buttons has been situated ergonomically close to the existing game buttons so they will be easy to press while driving. The demonstration application is able to record signals from the buttons in the log file which is consisting successful button clicks together with the time stamp in milliseconds. The log file data is currently in textual format so later in the project all the information will be stored in the database in the back-end system.

Backend system

The driving inspection and the future training environment systems have been designed to benefit from an automatized background data gathering system for purposes of research supporting analytics, personal progress tracking, and future opportunities in official or medical use. The data gathering system was created around the concept where the end user could track individual inspection cases with a high level of detail, allowing the test organizer (e.g. doctor) and test subject to play back replays of driving evaluation tests or training rounds. The system has also been designed to record any significant events during the simulation, such as events prompted to the end user to react upon, events of hazards or possible hazards (e.g. near-collision situation), and trackable driving situation events such as changing lanes or reacting to traffic signs.

The background system is based on a database capable of storing large amounts of information, thus a modern cloud based solution was selected as a primary platform already in the preliminary phase of the system architecture design. The database was designed to collect personal background information, basing the system on the existing pen-and-paper method of background data collecting, and extending the requirements based on discussions with the researchers from the research team. This background data involves non-identifiable personal information (e.g. gender and age), medical information (e.g. medicine use, eyesight, sleeping habits) and video game playing experience and habits. The second portion of the database is structured to store detailed information of all the data linked to the personal information data, so the database is designed to host data collections of medical substances, drugs, disorders and other medical conditions. This structure of data will allow gathering the personal background data in systematic and normalized manner. The third portion of the database structure will host large datasets to store all the driving test and training recordings. With each recording, personal background condition fields can be attached, for example allowing the system to store data of how many hours the person has slept before the recording, or if there has been medicine or other substance used within the past 24 hours.

The personal background data entry is done digitally, for example by the test arranging personnel or the test subject, employing easy-to-use entry methods, for example with digital forms on a tablet computer. User interface and usability design will be employed to ensure a very low error level in the data entry process. The current two-step system of pen-and-paper entry, which is then stored on a computer system presents an extra step which can act as an entry point for human error in systematic data gathering, and the newly designed digital system is based on the concept of minimizing the effect of human error in any data entry steps.

The driving simulation recording system is fully automated, and after the background data entry steps, the person attending the simulation will need no extra effort or attention for the data gathering to be executed. This recording system works in similar manner as in driving video games, where the whole session is recorded in the working memory of the computer, allowing instant playback of the simulation session, and furthermore storing the session data into the analytics database. The gathered session data structure is designed to host virtually all possible conditions and events happening during the simulation, thus allowing analytics in the future to research any circumstances that might have occurred within the simulation. Consequentially, this recording and playback engine allows researchers in the future to tap into vast datasets of driving simulation data gathered, and allows research on various topics, which might not even have been identified yet today. This form of vast datasets, or big data, creates a system that in the future will allow data analytics to find new patterns of data, for the purpose of medical research, official and regulatory use (e.g. police or traffic officials), and the data itself will become a valuable resource, which creates additional benefit and added value upon the training and test benefits of the simulation system.

Backend data security considerations

When storing sensitive personal data, such as medical information or substance use habits, even a local database would need a firm scrutiny of data security procedures involved within the whole system. When employing a cloud based server system for hosting the background and analytics database, the data security questions arise to a whole another level. Not only the database itself needs to follow strict data security guidelines, but all the data traffic between the simulation and background data gathering clients and the cloud servers need to be secure and encrypted.

In the preliminary data architecture design phase, discussions with the current stage of the cloud system data security were conducted, and the conclusion was that as long as personally identifiable data is not stored into the database, the data security level of current cloud systems is adequate for the background and analytics data gathering of the designed system. Furthermore, the reliable and big cloud system operators host their server computers in data centers, which have strict data security protocols, up to the level of having guards and restricted entry to the facilities itself. The digital data security level of the modern cloud systems most often is of higher level than an individual research organization could provide, with 24/7 data security support and administration provided by the cloud operator.

For future needs, the personally identifiable data, such as name or social security number can be hosted in another database, of higher level of security, and the identifier of personal information will be tokenized within the analytics and background data database. This tokenization ensures, that personal information can be revealed only one-way, and with officials (e.g. medical doctor or police) who have access to the personally identifiable data database, with higher level of data security policies. The personal information data cannot be revealed the other way around from the side of the background and simulation recording database, thus allowing a secure architecture to store sensitive data. This method of tokenization of personal identifiers will be designed during the research project, however for the prototype or proof-of-concept level system to be implemented during the project, the tokenization and personally identifiable data will not be stored into the database. This allows manageable data security level to be implemented during the research project, but the design can be in the future expanded to a more strict data security level solution via the tokenization methodology.

2.2. Participants

The participants were right-handed young and elderly adults. The young adult group consisted of 25 individuals (17-45 years of age; 14 female), and elderly group consisted of 20 (47-96 years of age; 11 female). The right-handedness was chosen here as an inclusion criterion, because our previous findings on spatial perceptual/attentional bias is based in right-handed population implicating a certain type of lateralization of brain functions [4]. Seniors mostly were recruited from foster (old-age, nursing) homes and elderly recreational clubs/facilities, whereas the young adults were mostly university students.

The participant had to be physically healthy with no neurological or psychiatric problems. Corrected vision and hearing were approved.

3. Procedure

3.1. Questionaires

The tests of the study partly conducted at the Centre for Cognitive Neuroscience, University of Turku. For the convenience of the elderly, the tests when necessary were conducted in the natural environment of the participants like recreational locations and old-age homes. Before the tests three forms were to be filled by the participant. First one was targeted usage of drugs and alcohol, hours of sleep, and level of alertness/tiredness at the time of testing. Also issues such as profession, driving experience and accidents, the experience with computer games etc. were asked. The second inquiry evaluated handedness. Third one focused more thoroughly on participants' game habits, and was filled only if the participant actively played computer games.

3.2. Driving Tasks

The driving task was adjusted (via a long series of piloting with drivers of different ages) to mimic driving on a normal countryside two-lane road with hills and curvatures and with a speed of about 100 km/h, and was considered rather easy for a person with usual driving experience. The most problematic part in driving task was the steering wheel - most manufacturers produced steering wheels that mimicked those for racing cars (F1), and thus were absolutely too different from the steering wheels of ordinary cars what comes to the feel of the steering wheel and the driving itself. After a long series of trials we ended with the steering wheel by Fanatec Porsche 911 GT3 RS V2 Wheel US (TM). The steering experience must be as close to the ordinary, because otherwise the unfamiliarity of the steering itself takes over too much of the driver's attentional capacity. This also is a challenge to manufacturers of the steering wheels for games.

The errors in driving, i.e. crossing the lane border, time spent on or outside the lane border, and the correction of errors were all stored to allow many types of error analyses. Here we only report number and side of crossing the lane border during the session.

3.3. Stimuli and measures of performance

The visual stimuli were white dots appearing in the periphery of the visual view in the lower left and right quadrants of the visual field (see Fig. 2). The location of the dots was constant (but can be varied for instance in applications for patients with blind spots in their visual field).

The auditory stimuli, applied via headphones (SHURE/SRH440) were short sine wave bursts of 550 Hz frequency and 50 ms duration, and enveloped to avoid the overshoot (clicks) in the beginning and the end of the burst. The intensity was adjusted to be 66 dB, and adjusted for each participant individually whenever necessary.

The auditory and visual stimuli were applied unilaterally to the left, right, and bilaterally in randomized order, and the participant's task was to react to the stimuli as quickly as possible by pressing corresponding button attached to the left and right sides of the steering wheel. Three different difficulty levels, high (ISI 500-1000 ms varying in steps of 13 ms), medium (ISI 700-1200 ms), and low (ISI 1000-2000 ms) were chosen after piloting of the system with participants of different ages. Some of the young participants showed 100% correct responses, i.e. ceiling effect, at the lowest difficulty level, whereas some of the old participants had zero correct responses at the difficult level. The correctness of the response (occurring within 150-1000 ms after the stmulus inset) and reaction times (RTs), i.e. speed of the correct responses, were determined as performance indicators.

3.4. Training and test sessions

The test sessions were always preceded by pilot runs, where the participants first got acquainted with the driving task and the steering wheel (2 min). Then a training session with just the visual and auditory stimuli and responding to them by button presses but without driving task (2 min) was accomplished. Finally, the driving task was combined with the stimuli and button presses (2 min).

After the training the test sessions were run each of the difficulty levels twice in counterbalanced order (ABC-CBA). The order of sessions was randomized for each participant. The duration of each test session was adjusted to be 2 min, which resulted in average numbers of both auditory and visual stimuli (uni- and bilateral together) being 24-25 in the fast (ISI 500-1000 ms), 18-20 in the medium (ISI 700-1200 ms), and 12-13 in the slow (ISI 1000-2000 ms) stimulus spacing, on the average. The order of auditory and visual stimuli and stimulus conditions were randomized. The participant's key task was to keep the vehicle on the lane, but also to detect as many stimuli as possible.

4. Results

4.1. Driving errors: Crossing over the lane borders

Fig. 3 shows the driving errors, i.e. crossing over the lane borders to right or left by the two age groups at the three task difficulty levels. Three-way mixed ANOVA [Factors: driving error side (left, right), difficulty level (fast, medium, slow), age (young, elderly)] applied to average percentage of errors to the left and right side crossings, resulted in significant main effects for driving error side, F(1,41) = 6.73, p = 0.01. Participants crossed the left lane border more than the right (Fig. 3).

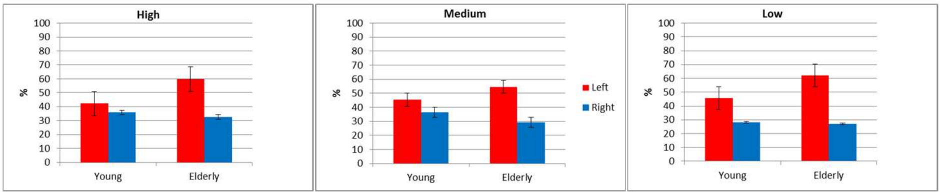


Figure 3. Driving errors (lane border crossing) by the two age groups at the three difficulty levels (spacing of the spatial stimuli), high(ISI=500-1000 ms), medium(ISI=700-1200 ms), and low(ISI=1000-2000 ms)

4.2. Auditory stimuli

Correct responses

Fig. 4 shows correct responses by the two age groups to three auditory stimulus conditions at three difficulty levels. Three-way mixed ANOVA [Factors: difficulty level (high, medium, low)* stimulus condition (left, bilateral, right)*age (Young, elderly)] revealed significant main effects for stimulus condition, F(1,44) = 30.89, p < 0.01, difficulty level F(1,61) = 174.30, p < 0.001, and age, F(1,41) = 10.42, p < 0.01. The young in general outperformed the elderly. There was a significant interaction between difficulty level and stimulus condition, F(3,108) = 4.10, p = 0.01. The effects were further investigated with pairwise comparisons using Fisher's LSD tests. Participants gave a smaller number of correct responses at the high (M = 40) than at the medium (M = 61) and at the low (M = 75) difficulty levels [high-medium (p < 0.01), high-low (p < 0.01), medium-low (p = 0.01)]. At the high difficulty level, the participants gave more correct responses to the right (M = 47) than to the left (M = 44) stimuli (P < 0.01). At the medium difficulty level, more correct responses were given to the right (M = 70) than to the bilateral (M = 45) stimuli (p < 0.01). At the low difficulty level, more correct responses were given to the left (M = 83) than to the bilateral (M = 58) stimuli (p < 0.01) (Fig. 4, bottom).

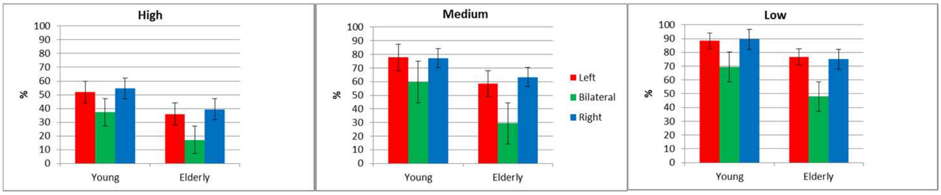


Figure 4. Correct responses in the three auditory stimulus conditions, left, bilateral, and right, by the two age groups at the three difficulty levels, High(ISI=500-1000 ms), medium(ISI=700-1200 ms), and low(ISI=1000-2000 ms).

RTs

Figure 5 shows the RTs of the two age groups at the three difficulty levels in the three auditory stimulus conditions. Three-way mixed ANOVA [Factors: difficulty level (high, medium, low)*stimulus condition (left, bilateral, right)*age (young, elderly)] revealed significant main effects for difficulty level, F(2,68) = 58.16, p < 0.01, stimulus condition, F(2,50) = 43.21, p < 0.01, and age, F(1,34) = 6.90, p = 0.01. The elderly in general were slower than the young. An interaction between difficulty level and stimulus condition, F(3,86) = 13.36, p < 0.01, was detected. Further analysis using Fisher's LSD pairwise comparisons was conducted. Significantly longer RTs were obtained at the high (M = 570) difficulty level followed by low (M = 663) and medium (M = 635) difficulty levels [high-medium (p < 0.01), high-low (p < 0.01), medium low (p < 0.01)]. At the medium difficulty level, the participants reacted faster to the right (M = 603) than to the left (M = 626) (P < 0.01) stimuli (Fig. 5, middle). At the low difficulty level, the participants reacted faster to the right (M = 619) than to the left (M = 633) (p = 0. 02) or than to the bilateral (M = 731) (p < 0.01) stimuli (Fig. 5, bottom).

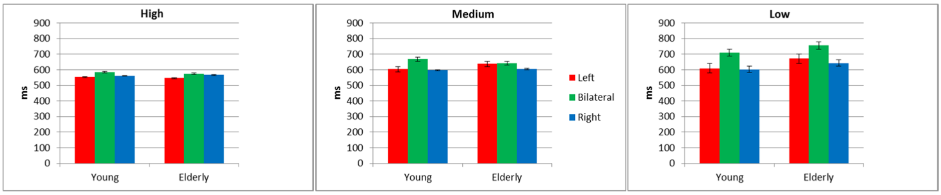


Figure 5. Reaction times in the three auditory stimulus conditions, left, bilateral, and right, by each group at the three difficulty levels, high(ISI=500-1000 ms), medium(ISI=700-1200 ms), and low(ISI=1000-2000 ms).

Erroneous unilateral responses to bilateral stimuli

Fig. 6 shows erroneous unilateral responses in the bilateral stimulus condition by the two age groups at the three difficulty levels. Three-way mixed ANOVA [Factors: difficulty level (high, medium, low)* response side (left, right)*age (young, elderly)] revealed significant effect for difficulty level, F(2,70) = 10.93, p < 0.01, and an interaction between difficulty level, response side, and age, F(2, 82) = 8.15, p < 0.01. Further analysis with pairwise comparisons using Fisher's LSD indicated that the largest number of errors occurred at the high (M = 47) difficulty level and the smallest at the low (M = 38) difficulty level [high-low (p < 0.01); medium-low (p < 0.01)]. At the low difficulty level, the elderly responded more correctly (M = 43) than the young (M = 32) (p = 0.03). At the high difficulty level, the young responded more to the left (M = 58) than to the right (M = 34), paired sample t(24) = 2.82, P = 0.01, whereas elderly responded more to the right (M = 59) than to the left (M = 38), though not significantly, paired sample t(19) = -1.95, p = 0.06. Independent t-tests showed that the young and the elderly differed in their left responses (M = 59, M = 38, respectively), t(43) = -2.57, p = 0.01, and also in their right side responses(M = 34, M = 58, respectively), t(43) = 3.63, p < 0.01.

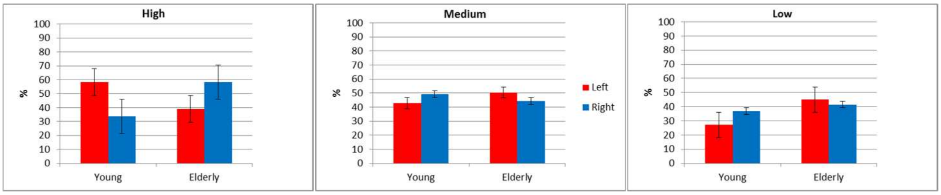


Figure 6. Erroneous unilateral responses to the bilateral auditory stimuli by the two age groups at the three difficulty levels, high(ISI=500-1000 ms), medium(ISI=700-1200 ms), and low(ISI=1000-2000 ms).

4.3. Visual stimuli

Correct responses

Fig. 7 shows correct responses of the two age groups at the three difficulty levels in the three visual stimulus conditions. A three-way mixed ANOVA [Factors: stimulus condition(left, bilateral, right)*difficulty level(high, medium, low)*age(young, elderly)] revealed significant main effects for difficulty level, F(2,58) = 92.45, p < 0.01; stimulus condition, F(2,67) = 7.37, p < 0.01; and age F(1,40) = 21.09, p < 0.01. The young outperformed the elderly. Pairwise comparisons of the difficulty levels using Fisher's LSD tests showed that the smallest number of correct responses were obtained at high (M = 46) difficulty level, more at the medium (M = 63) and the largest number at the low difficulty level (M = 75) [high-medium (p < 0.01), high-low (p < 0.01), medium-low (P < 0.01)]. Pairwise comparisons of the stimulus conditions showed that participants gave more correct responses to the right (M = 63) than to the left (M = 61) (P < 0.01) or than to the bilateral stimuli (M = 59) (p < 0.01).

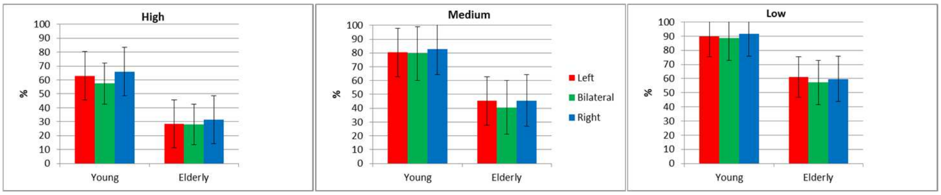


Figure 7. Correct responses to the three visual stimulus conditions, left, bilateral, and right, by the two age groups at the three difficulty levels, high(ISI=500-1000 ms), medium(ISI=700-1200 ms), and low(ISI=1000-2000 ms).

RTs

Fig. 8 shows the RTs at the three difficulty levels in the three stimulus conditions by the two age groups. Three-way mixed ANOVA [Factors: stimulus condition (left, bilateral, right)*difficulty level (high, medium, low)*age(young, elderly)] revealed significant main effects for difficulty level, F(2.72) = 12.24, p < 0.01, stimulus condition F(2,72) = 11.39, p < 0.01; and age, F(1,36) = 16.24, p < 0.01. The young reacted faster than the elderly.

Pairwise comparisons wish Fisher's LSD tests showed that the RTs were shortest at the high (M = 511), longer at the medium (M = 537) and longest at the low (M = 546) difficulty levels [high-low (p < 0.01), high-medium (p < 0.01)]. Pairwise comparisons of the stimulus conditions showed that participants reacted faster to the right (M = 515) than to the left (M = 537) (P < 0.01) or than to the bilateral stimuli (M = 541) (p < 0.01).

.

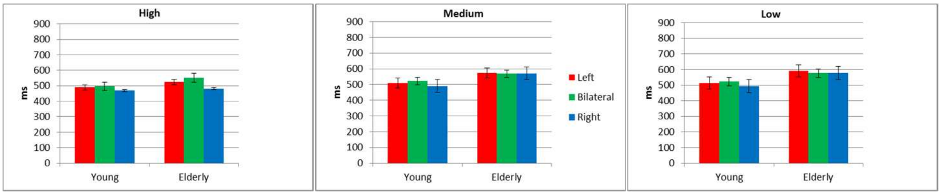


Figure 8. Reaction times to the three visual stimulus conditions, left, bilateral, and right, by the two age groups at the three difficulty levels, high(ISI=500-1000 ms), medium(ISI=700-1200 ms), and low(ISI=1000-2000 ms).

Erroneous unilateral responses to bilateral stimuli

Fig. 9 shows erroneous unilateral responses of the age groups to three stimulus conditions at three difficulty levels. Three-way mixed ANOVA [Factors: difficulty level (high, medium, low)*response side (left, right)*age (young, elderly)] showed significant main effects for difficulty level, F(2,75) = 40.35, p < 0.01, response side, F(1,40) = 16.45, p < 0.01, and age, F(1,40) = 6.80, p = 01. In general, more responses were made to the right than to the left. Elderly made more errors than young. Pairwise comparisons of the difficulty levels using Fisher's LSD showed that the largest number of erroneous responses were given at the high (M = 49) difficulty level, a smaller number at the low (M = 24) and the smallest at the medium (M = 41) difficulty levels [high-medium (P < 0.1), high-slow (P < 0.1), medium-slow (P < 0.1)].

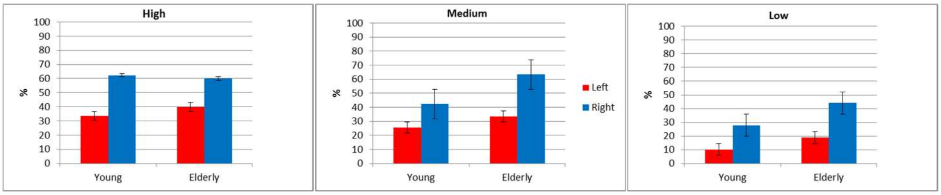


Figure 9. Erroneous unilateral responses to the bilateral stimuli by the two age groups at the three difficulty levels, high(ISI=500-1000 ms), medium(ISI=700-1200 ms), and low(ISI=1000-2000 ms).

5. Discussion

The results demonstrated the rather equal performance level of the two right-handed age groups in the virtual driving task, and the preference of the right side stimuli (significant for auditory) especially by the elderly participants. Also a slight superiority of the young over the elderly participants in their perceptual/attentional performance level while driving was obtained. The most unexpected finding was the simultaneous occurrence of rightward bias in spatial perception and attention and the leftward bias in driving errors. This was evident in both age groups but slightly more prominent in the elderly. The expected finding was the rightward bias in the perceptual and attentional capacity, which was most evident under larger cognitive load. Distinct rightward bias was now found for both the visual and the auditory spatial stimuli (cf. [2, 3] where the phenomenon was more distinct for auditory than visual stimuli).

A curious feature in driving was the bias in lane border crossings to the left. We have to focus on this phenomenon when more data is gathered. If it is true and more specifically, if it proves to be a stronger phenomenon among the older drivers, then the questions rises whether it has something to do with right-handedness and the spatial perceptual-attentional bias to the right by our present participants. We have originally suggested that the rightward spatial perceptual/attentional bias seen in our previous laboratory measurements [2, 3] is due to the functional lateralization of the normal and especially normally aging brain [4]. Furthermore, it has been shown previously that patients with hemispatial neglect, the pathological rightward bias of perception and attention due to injury of the right hemisphere attention mechanisms, have a bias to contralesional (left) collisions [6, 7] when navigating a powered wheelchair. This is in contrast with the same patients in [7] study showing a rightward bias when walking, which reveals the task dependency of the spatial bias. Indeed, our preliminary findings indicate that there may be a connection between rightward bias in spatial perception and attention and the leftward bias in steering a vehicle in the normal and normally aging right-handed population. The positioning of the vehicle on the road (on the lane, between other vehicles) has been shown to be one of the key skills declining with aging [8, 9]. These real-world phenomena can be reliably demonstrated even with a simple virtual system introduced in the present study.

The right-side bias found with the virtual system was expected since the phenomenon has been demonstrated in previous laboratory studies with both auditory [2] and visual [3] linguistic and nonlinguistic stimuli. In contrast to those previous studies, in the present results the bias in visual modality was an even more distinct that that in the auditory modality. This phenomenon needs further attention while more data, and possibly under driving tasks with different demands, is gathered. If it proves to be real it may indicate that the visuomotor driving task may induce changes in the processing of visual information, i.e. both tasks load the same visuo-spatial system, thus resulting in more bias in visual spatial perception.

The simultaneous driving/spatial perception/attention task works well. Not many (elderly) participants interrupted the task, and also errors were obtained in both age groups. As revealed by the performance of the young and elderly at lowest and highest difficulty levels, we feel that it is justified and necessary to have the different difficulty levels in the tasks. Because of the varying number of each stimulus modality and condition within the 2 min session, we already have modified the system into one, where the exact number of each stimulus condition is predetermined by the experimenter, and the session duration is at least 2 min. As an alternative one could also think of a system which begins at low difficulty level, but then increases the difficulty level continuously until a clear reduction in performance level (e.g., 75%) is reached.

The present low-cost system is designed for testing and evaluation purposes, but it can be used also as a template for training programs in the future. In the future versions also rear-view mirrors, pedals (plus gears), and turn-lights will be included in the system.

The relative amount of elderly population is growing at accelerating speed, and simultaneously an outstanding growth is seen in the traffic. This produces increasing demands to the driving skills and overall cognitive perceptual and attentional skills of the aging drivers, and thus means for extra training in safe environments must be developed. Recently very promising results have been reported related to the effects by training in the virtual environments on the working memory and executive functions (see e.g. [10, 11] which indeed are in focus in traffic behavior.

Acknowledgement

We would like to thank the funding agency, Finnish Funding Agency for Technology and Innovation (Tekes), The Traffic Safety Committee of Insurance Companies (VALT), Henry Ford Foundation, our academic and industrial partners, and the experts from the Turku Game Lab (TUAS) Finland. We also thank all the test subjects who took part in the experiments.

References

[1] Takio F, Koivisto M, Jokiranta L, Rashid F, Kallio J, Tuominen T, Laukka S J, Hämäläinen H: The effect of age on attentional modulation in dichotic listening. Developmental Neuropsychology, Vol, 34, pp. 225–239, 2009 DOI:10.1080/87565640902805669

[2] Takio F, Koivisto M, Laukka S J., & Hämäläinen H: Auditory rightward spatial bias varies as a function of age. Developmental Neuropsychology, Vol. 36, pp. 367- 387, 2011. DOI:10.1080/87565641.2010.549984

[3] Takio F, Koivisto M, Tuominen T, Laukka S J, Hämäläinen H: Visual rightward spatial bias varies as a function of age. Laterality: Asymmetries of Body, Brain and Cognition, e-pub, 1-24, 2012 DOI:10.1080/1357650X.2011.628675

[4] Takio F, Koivisto, M, & Hämäläinen, H: The influence of executive functions on spatial biases varies during the lifespan. Developmental Cognitive Neuroscience, Vol. 10, 170–180, 2014. DOI:10.1016/j.dcn.2014.09.004

[5] Luimula M, Besz A, Pitkäkangas P, Suominen T, Smed J, Izullah, FR, and Hämäläinen H: Virtual Evaluation Tool in Driving Inspection and Training. in Proceedings of the 5th IEEE Conference on Cognitive Infocommunications, Gyor, Hungary, pp. 57-60, 2015 DOI: 10.1109/CogInfoCom.2015.7390564

[6] Punt TD, Kitadono K, Hulleman J, Humphreys, GW, Riddoch, MJ: From both sides now: crossover effects influence navigation in patients with unilateral neglect. Journal of Neurology, Neurosurgery, and Psychiatry, Vol. 79, pp. 464-466, 2008 DOI:10.1136/jnnp.2007.129205

[7] Turton A J, Dewar SJ, Lievesley A, O'Leary K, Gabb J, & Gilchrist, I D: Walking and wheelchair navigation in patients with left visual neglect. Neuropsychological Rehabilitation, Vol. 19, pp. 274-290, 2009 DOI: 0.1080/09602010802106478

[8] Di Stefano M, MacDonald W: Assessment of older drivers: Relationships among on-road errors, medical conditions, and test outcomes. Journal of Safety Research, Vol. 34, pp. 415–429, 2003 DOI:10.1016/j.jsr.2003.09.010

[9] Baldock MRJ, Berndt A, Mathias, JL: The functional correlates of older drivers' on-road driving test errors. Topics in Geriatric Rehabilitation, Vol. 24, pp. 204-223, 2008 DOI:10.1097/01.TGR.0000333754.90550.b5

[10] Anguera JA, Boccanfuso J, Rintoul JL, Al-Hashimi O, Faraji F, Janowich J, Gazzaley A: Video game training enhances cognitive control in older adults. Nature, 5 01, pp. 97–101, 2013 DOI: 10.1038/nature12486

[11] Casutt G, Theill N, Martin M, Keller M, Jäncke L: The drive-wise project: driving simulator training increases real driving performance in healthy older drivers. Frontiers in Aging Neuroscience 6:85, 2014. DOI:10.3389/fnagi.2014.00085

Acta
Technica
Jaurinensis

Vol. 10, No. 1, pp. 35-49, 2017
DOI: 10.14513/actatechjaur.v10.n1.412
Available online at acta.sze.hu

Gamified Solutions in Healthcare - Testing Rehabilitation Games in Finland and Asia

C. Kattimeri[1], P. Qvist[1], N. Katajapuu[1], P. Pitkäkangas[1], H. Malmivirta[1], M. Luimula[1], A. Pyae[2], T.N. Liukkonen[2], and J. Smed[2]

[1]Turku University of Applied Sciences, Turku Game Lab
Joukahaisenkatu 3, 20520 Turku, Finland
Phone: +358 403550839
e-mail: mika.luimula@turkuamk.fi

[2]University of Turku, Department of Information Technology
FI-20014 University of Turku, Finland

Abstract: This paper presents a comprehensive summary of the Gamified Solutions in Healthcare (GSH) research project, which is a joint research project between Turku University of Applied Sciences and the University of Turku. The goal of the project is to promote exercise, social inclusiveness and enhance quality of life, aiming at developing new services and effective activity solutions for the elderly through gamification. During the research project elderly people were included in the development and testing of games so that they could be used for more than just entertainment purposes. According to all of our tests elderly enjoy playing exergames, and digital games can be an effective way to enhance the quality of life of the elderly. In the case studies it was observed that the players where motivated while playing but motivation should also be maintained throughout the gameplay. The elderly gave overall positive feedback for the idea of using digital activity games for exercising.

Keywords: Gamification, Serious Games, Rehabilitation, Usability, Field Experiments

1. Introduction

According to the World Health Organization active ageing is the process of optimizing opportunities for health, participation and security in order to enhance quality of life as people age. Seniors are easily alienated from the society due to the digital divide and unequal opportunities of using modern technology; therefore it is very important to include them in the development of the information society [1].

In the industrialized and developed countries, the growth of both number and proportion of the elderly in the population is evident, and the elderly is one of the most rapidly growing age groups [2]. The age-related decline of physical and cognitive capabilities for an individual can present significant impact on the quality of life and furthermore elevate the need for additional social and health related services. These two factors combined can lead to rising healthcare and wellbeing related costs on the individual and societal levels. Providing physical exercise to the muscles [3] and cognitive and memory related exercise to the brains [4] can alleviate, slow down, reverse and even stop this decline [5, 6].

Nowadays, the use of digital games is not only restricted for the purpose of entertainment, but expanded in areas such as education, healthcare, business and the military as well. Digital games are designed to make the players experience high levels of motivation and engagement in the game itself. One of the key motivators for starting to further study the gamified methods to apply for healthcare was to employ this high motivational level and engagement found within digital games in useful and productive application areas. These games designed for other primary purposes than solely for entertainment purpose are generally defined as serious games. Furthermore, gamification is the term used for the process of applying game-like thinking and game mechanics to traditionally non-game applications and functions (e.g. education and exercise) to make them more fun and engaging [1].

This paper is structured to present the Gamified Solutions in Healthcare research project, and consequentially the serious games developed within this research project, and furthermore discussing the findings and results of various studies and tests conducted during the timeframe of the research project. The developed games and the relevant tests are presented as case studies, and a literature review on motivational factors for elderly stroke patients is also included. The paper concludes with the summary of various tests conducted during the research project, presenting a summary of the impact of these tests and a compilation of test groups with the number of test subjects involved in the studies from all the countries participated in the tests of the research project.

2. Gamified Solutions in Healthcare

Gamified Solutions in Healthcare (GSH) is a joint research project between Turku University of Applied Sciences and the University of Turku. In the course of this project, new gamified services are researched and developed in cooperation with Serious Games Finland Oy, Attendo Finland Oy, City of Turku Welfare Division and Puuha Group Oy. The goal of the project is to promote exercise, social inclusiveness and enhance quality of life. GSH is funded by Tekes, the Finnish Funding Agency for Innovation [1].

The GSH project aims at developing new services and effective activity solutions for the elderly through gamification. The aim of the GSH project is to include elderly people in the development and testing of games so that they could be used for more than just entertainment purposes.

In the course of the GSH project we will conduct more research about the attitudes of the seniors towards game-playing and digital games in general. We need more focus group interviews as well as some usability testing with console games. The research requires systematic analysis about the existing games for seniors, and the attitudes and perceptions of health and social care workers that work with seniors. In summary, our research topics include: gamification mechanisms, usability for elderly people, the effectiveness of gamified solutions for elderly people (e.g. business and production models), and attitudes and acceptance of games by the elderly people [1].

The Finnish game development industry needs new solutions for active health ageing and, therefore, cooperation with Asian partners will open up new possibilities to build innovations and generate new business opportunities. Researchers around the world are now activating even if this research field is still quite new [1]. Dealing with the ageing population is a challenge not only in Asian societies but also in Western societies such as Finland. Therefore, research organizations are very active in the field of serious game development and developing games for elderly. Thus, we believe that these kinds of rehabilitation concepts will be needed in the future [7, 8].

One of the initial goals of the GSH project was to design and implement a Virtual Nursing Home (VNH) in collaboration with our industrial partners Puuha Group, GoodLife Technology, City of Turku and Attendo Oy in Finland. VNH is a concept that provides alternative solutions for the elderly's self-management and self-care, and prevents them from social isolation while, at the same time, reduces the workload of the healthcare professionals. The methodologies used for the implementation of this project were User-Centered Design (UCD) and rapid prototyping [9].

3. Case Studies in GSH Project

In this section we present six studies we have conducted related to the usability of gamification in encouraging physical or mental activity. Although most of these studies take different approaches and have different test groups, they reveal features that are common to physical activity oriented gamification. In addition, we present a literature review on motivational factors for elderly stroke patients.

3.1. Pre-study: Testing Glider game in Japan

In 2013, we focused on the usability evaluation of a prototype game called "The Glider", which is a Kinect-based game by Serious Games Finland (SGF). The study identifies difficulties, and suggests improvements related to the user interface development of "The Glider". The controlling motions in the game have been developed in close cooperation with medical doctors, physiotherapists, and patients in the terms of well-being and light exercise. This study utilizes methods based on Kansei Engineering. The Japanese word "Kansei" can be interpreted as emotion, feeling, receptivity, or sensibility. In essence, Kansei Engineering is a consumer-oriented technological research field that supports engineers in the development or improvement of products via evaluating the "Kansei" of the users [7, 8].

We conducted usability tests with 12 test subjects (6 male and 6 female) who were all Japanese students. The test subjects played the prototype game and spoke aloud their thoughts while playing. Two video cameras were recording the test scene [7, 8].

The test results showed that the beginning of the game was very simple and easy to play having just one motion control per level. The players were learning the exercises while playing the first levels; however, they should receive more information and guidance before starting to play. Some of the motion controls that were developed for rehabilitation increased the complexity of the game. The players were motivated while playing the game. Moreover, difficulty levels and the understanding of the game seemed to be in balance. The players were interested to play and learn new things; however, after the first four levels they seemed to slightly lose their motivation. In these four levels, the players were iteratively reminded of how to control the glider (accelerating, and moving in all directions). This is needed later with more complex steering motions, although the next levels showed that the players were not able to control the left and right movement anymore. Generally speaking, the game flow seemed to be in balance between difficulty and usability from the beginning until the fourth level [7, 8].

3.2. Case SportWall and commercial games

We investigated the usability of two commercial games and a physical activity game developed by our project collaborator, Puuha Group, Finland. The main objectives of this pilot study were: to investigate the usability and usefulness of games for the elderly, to evaluate the usability and usefulness of multimodal input devices (Kinect, PlayMove, web-camera/Xtreme reality) for the elderly, and to understand the general user experiences of elderly in playing games [1, 9, 10].

The games chosen for the tests were Puuha Group's SportWall (using Xtreme reality technology and a web-camera), Microsoft Xbox's Kinect-based climbing game and Playstation3's PlayMove tennis game. In the SportWall game the player uses particular body postures and gestures to control the roller-skating character in the game [1, 9, 10].

Our test participants were two groups of elderly between the ages of 65-85 years. The first group had 5 moderately active elderly, whereas the second group had 5 less active elderly but only 2 showed up for the tests. The second group had health problems and limited short-term memory and they could not participate in the tests or answer the questionnaires. Consequently, we only collected feedback from the moderately active elderly group and our results are based on that. We used two video cameras (front and back views) to record the actions and gameplay of the elderly. In addition we carried out questionnaire sessions that were captured by a voice recorder [1, 9, 10].

In general, the elderly participants in our study were somewhat physically active and they did not have prior experience with digital games. In the test of the climbing game, the elderly had some problems in the beginning but after the tutorial and the guidance from the researchers they could manage without major challenges. Also, we found out that the user-interface, music and audio feedback in the game were not elderly-friendly. However, Kinect-based interaction was effective for the elderly. In the test of the tennis game, the elderly faced challenges with the buttons of the PlayMove controller. Nevertheless, they liked the idea of the tennis game. According to the test participants the last game, SportWall, was simple and had a clean interface. The interaction with the game was easy due to the use of a traditional webcam. However, some game actions (e.g. jump, sit) were unsafe for the elderly. In general, Kinect for the Xbox One was the most effective input device for the elderly yet the scores of the other two devices were not noticeably different [1, 9, 10].

According to the interview with the caregiver of the service home, game-based physical activities are interesting and they can improve the motivation of the elderly in doing physical exercises. The socialization of the elderly can also be enhanced. The caregiver advised that the games should be simple and easy because of the elderly's limitations in mobility and memory [1, 9, 10].

Furthermore, we conducted a pre-study that consisted of gathering and studying existing games for seniors, conducting a pre-test on console games and interviewing potential users of serious games. The search was limited to games that could enhance physical, mental or social well-being. In addition, the senior-friendly or unfriendly features of the discovered games were evaluated. Also, the barriers or attitudes of introducing games for the elderly were examined. The findings from the pre-studies revealed the limitations of the existing games and technologies. Not all the commercial games are accessible for the elderly. Additionally, most exergames do not support elderly-friendly design and gameplay. The limitations of the current Kinect-based exercise games are lack of game customization, lack of effective feedback, and lack of long-term study. However, it was discovered that the existing games have potential in being reused with further modifications and enhancements in the future game development and testing within the GSH project. In general, the area of gamification for elderly and healthcare has noticeable challenges that need to be faced so that the researchers can investigate the benefits of digital games for elderly if any exist [10].

3.3. Case Old Photos on Map

The population in Europe is ageing dramatically and the ageing situation in Finland is even more challenging. Therefore, we propose "Old Photos on Map" (Vanhat kuvat) web-application to activate the memory of the ageing population [11]. Malmivirta et al. have found that several researchers have come to the conclusion that cognitive mental activity is of high-importance for mental health. The effectiveness of using old photos to enhance brain health has been examined in the research and development project "Art and culture – Keys for better Brain Health" [12].

Malmivirta et al. advocate that in autobiographical activities like watching and sharing experiences of old photos, exploring punctum photographs allows an emotional connection to be formed with the most significant people, places and events in our lives. Subjective sense of time refers to the ability to shift to thinking about something that has happened in the past. Memory layers and visual perception are in active motion in this phase, when narratives, photographs and maps are integrated into each other. The autobiographical memory and discussions about the old photos require complicated cooperation between our cognitive and emotional processes which activate different parts of the brain. The brain feels well when the environment promotes activation and involves social interaction. The emergence of memory disorders can be delayed by keeping the brain occupied with new and complex tasks throughout our lives [11].

The aim of "Old Photos on Map" web application is to wake up peoples' memories and experiences from their childhood and earlier life experiences as well as personal hidden stories. The "Old Photos on Map" web application research and development work applied the practical action research strategy in which the main objective was to develop an application to activate cognitive memory health of the aged people. During the research and development process, any need for changes was recorded in line with the objectives and also the pedagogic plans with the interventions of the test groups.

During the research and development process, the tests were conducted in two ways. Firstly we examined the design issues related to the user interface of the application. A virtual agent in the application encourages discussions between people in a group. We compared multiple agents and interface designs with different color variations and the final interface elements were chosen through questioning 24 Finnish people between 56-93 years old. According to the results from the questionnaires larger text was preferable. The agent that was chosen was described as kind, adult, trustful and interesting [6].

Secondly, we studied the "Old Photos on Map" web application without the agent focusing more on the interventions with the group activities of the elderly people using means-ends art pedagogy and social cultural learning strategy. At the same time we are recording the changes that are needed for the application. We had three test settings with the collaboration of the third sector organizations. In the first two groups the participants did not have any diagnostic memory deceases (average age 30-75 years, n=7) and in the third group the members had all some kind of diagnosed memory problems (average age 55-80 years, n=14). The tests have been conducted during May and June 2016 and one more will be made on August 2016.

The first look at the data shows that the application "Old Photos on Map" activates the memory of the participants and brings out past memories. It also makes the participants to share their memories and learn more from each other`s memories. It seems that when looking at the old photos on the map, diving into the punctum of the old photos and exploring the layers of the photos it allows an emotional connection to be formed with the most significant places, events and people in our lives. It also seems that a special pedagogical frame is needed for the interventions with different kinds of people and groups for revealing the hidden stories of the photos used. The test results seem to be very promising also with the people suffering with memory problems. While looking at old photos on the map a lot of sharp memories came out to be shared. The overall activity was growing, the eyes were sharpening and social interaction was richer.

3.4. Case SportWall and PhysioWall

We studied the usability and reception regarding SportWall (Game A) and PhysioWall (Game B) using SUS and GEQ questionnaires and interviews. In total we had 19 participants with overall average age 72±11 years (n=19). We had two testing settings: urban and rural. In the urban setting we had seven participants (n=7, 3 female and 4 male), and in the second testing session in the rural setting we had 12 (n=12, 7 female and 5 male) participants. Age distribution was similar in both settings, and both genders were represented equally (10 female, 9 male) [13].

These games have been our first games that were tested with the participants that belong to our targeted age group. From these games we have learned what kind of things seem to work with our target group and what does not work. To remedy the problems we found during the testing, we are currently doing a redesign on both games. We had designed the games to have clear graphics without a screen clutter, which is seen on games aimed for younger players. Despite our efforts, the participants still had problems with knowing what they should avoid and what they could collect on the Game A. The graphics problem might be partially caused by the unfamiliarity of the activity represented in the game A as the interviews and comments by the elderlies during the gameplay revealed their confusion about these matters. The speed in which things happen on the Game A was also a problem, albeit the game was significantly slower than the original version which was designed for the younger players. The next version of the game will have its pacing further slowed down, and the context of the game will be more familiar to the target group [13].

Game B had originally a virtual instructor whose movements' the player should match, and a graph that showed what kind of movements will be coming after the current one. This caused confusion as some of the participants were not sure which one to follow or they missed the correct movements because the extra graph grabbed their attention. In Game B, players had to raise their hands several times above their heads. This was a problematic procedure for several elderly, and prevented them from completing the required movements. These problems are addressed in the new version of the game where e.g. the indicator for the upcoming motion has been removed [13].

In the future field tests, the gaming sessions will be longer as it was noted that during the short sessions the participants were aware about their surroundings and the observers. The longer sessions might help the participants to relax and immerse more on the game, instead of them being filling questionnaires in every turn after few minutes of the play. This might allow us to see more emotions rising from the gameplay experience [13].

SUS is a simple tool for measuring basic usability of a software-based artefact, and it proved to be usable and informative also with the elderly, mainly thanks to its brevity. GEQ in the other hand is a questionnaire which contains 50 or in the case of multiplayer situation 67 questions. The length of the GEQ brought up questions and comments from the participants during the testing sessions (e.g. "When will this questionnaire end?", "How many questions are left?"). In the future, the game experience research with elderly, or with other people who do not play games regularly or at all, would benefit from a shorter and more understandable questionnaire [13].

3.5. Case Skiing Game

We evaluated the usability of an exergame called "Skiing game" for Japanese elderly. We also investigated the elderly's engagement in playing the Skiing game. The study was conducted at a therapy training room at the Sendai City Health Promotion center in Japan. We recruited 24 elderly participants between the ages of 60 and 85 years. In the usability testing, the elderly participants played the Skiing game, which is a game- based Skiing activity, originally designed for the Finnish elderly. To play the game, the elderly participants moved their hands forward and backward as a conventional double-pole Skiing activity. A traditional webcam that uses Xtreme Reality technology was used to detect the player's movements in the game. A screenshot from the Skiing game can be seen in Figure 1. Before the participants started playing the game they were asked questions regarding their attitude towards physical exercise. After they have played the game they were asked questions about the usability of the game, their engagement in the gameplay, their motivation, and attitude towards game-based exercise [14].

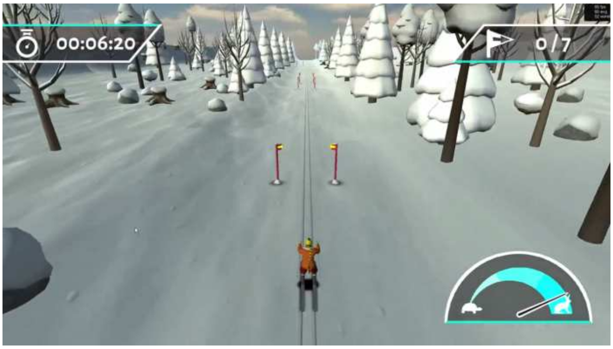


Figure 1 Screenshot from the Skiing Game

Based on the findings from the usability testing in Japan, we observed that the experiences and feedback of the elderly during and after the gameplay were noticeably positive. They were interested in playing the game and they would like to play it again. Therefore we claim that the Skiing game is an easy and user-friendly game for the Japanese elderly. Furthermore, most of the elderly agreed that playing digital games is an easy and effective way of exercising [14].

According to our observations from this study, we recommend the following design guidelines for usability practitioners. Firstly, we should take into consideration that the elderly should not be distracted from the gameplay therefore the game interface, context and contents should be simple and uncluttered. Secondly, excessive in-game instructions and unnecessary audio feedback should be avoided. Visual cues are important in the game, especially for novice elderly players. Controller-free and gesture-based interaction is effective for the elderly. Additionally, natural and familiar game actions can engage the elderly players more. To reduce the elderly's frustration while playing is always important, even if they did not achieve a particular task. Finally, while designing a game for elderly, it is very important to reduce and prevent the risk of falling [14].

In the course of the GSH project we tested and evaluated the usability of the Skiing game for Finnish elderly as well. The study was conducted at an elderly service home in Finland. We recruited 21 elderly participants, who were aged between 60 and 85 years. The results of the Finnish usability tests as well as a comparison between the Japanese and Finnish tests will be published in the near future.

3.6. Case Brain Trainer Exercise Game

This study was focusing on verifying the level (+ or -) of brain activity during the test subjects playing the "Brain Trainer Exercise Game" developed by Sendai Television, and furthermore evaluating the usability and acceptability of the game by observing the test subject behaviour during the gameplay. The test groups comprised of test subjects from Finnish and Japanese elderly population [15].

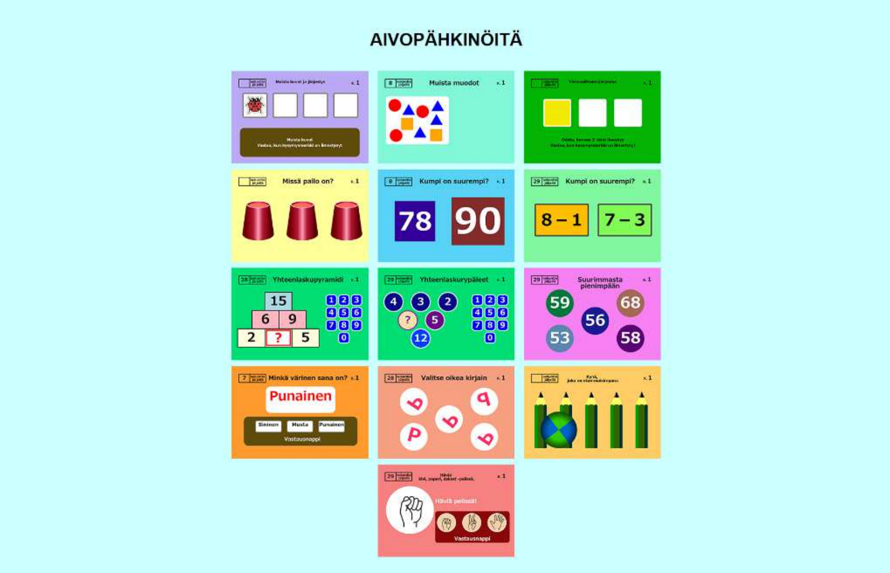


Figure 2 Brain Trainer Exercise Games main menu localised in Finnish language showing all 13 minigames.

The Brain Trainer Exercise Game was developed for the purpose of enhancing the cognitive skills of elderly population. The game includes 13 minigames shown in Figure 2 and one pen and paper calculation task. The main menu allows the player to freely choose their preferred minigame in any desired order. Each minigame has four different difficulty levels progressing from easiest (*) towards the most difficult (****), with the difficulty being derived from different parameters depending on the theme of the minigame. These varying difficulty parameters can consist of for example limited time for the task within the minigame or number of objects appearing during the minigame for cognitive tasks such as memory exercise. Varying the difficulty parameters allows the game to provide more challenging tasks for cognitive skills related to perception or requiring more speed from the player to provide the correct answers to the given tasks. The thirteen minigames focus on different areas of cognitive and memory skills, with themes ranging from numeric calculations to remembering different colors and shapes, and following the moving objects and afterwards remembering different tasks. The Brain Trainer Exercise Game was designed to be played with easy-to-use user interface with a device that employs a touch screen. The game was localised to Finnish language by Turku University of Applied Sciences [15].

During the study the field tests were conducted in Turku, Finland, by Turku University of Applied Sciences in December 2015, and in Sendai, Japan, by Sendai City Industrial Promotion Organization (Sendai-Finland Wellbeing Center, Business Innovation International Unit) in February 2016. The field tests consisted of three different test setups, with the first one utilizing a near infra-red spectroscopy (NIRS) device to monitor the brain activity level of the test subject during the gameplay. The second test setup involved the test subjects playing the game in three groups of 4-5 people, and the last test setup allowed the test subjects to play the game individually. Within the second and third test setup the usability and acceptability of the Brain Trainer Exercise Game was evaluated with questionnaires and interviews developed by Sendai Television Broadcasting Company. These questionnaires were translated into Finnish for the Finnish test groups. All the different test setups also involved direct observation of the test subjects. The tests were conducted according to the same exactly protocol in Finland and Japan [15].

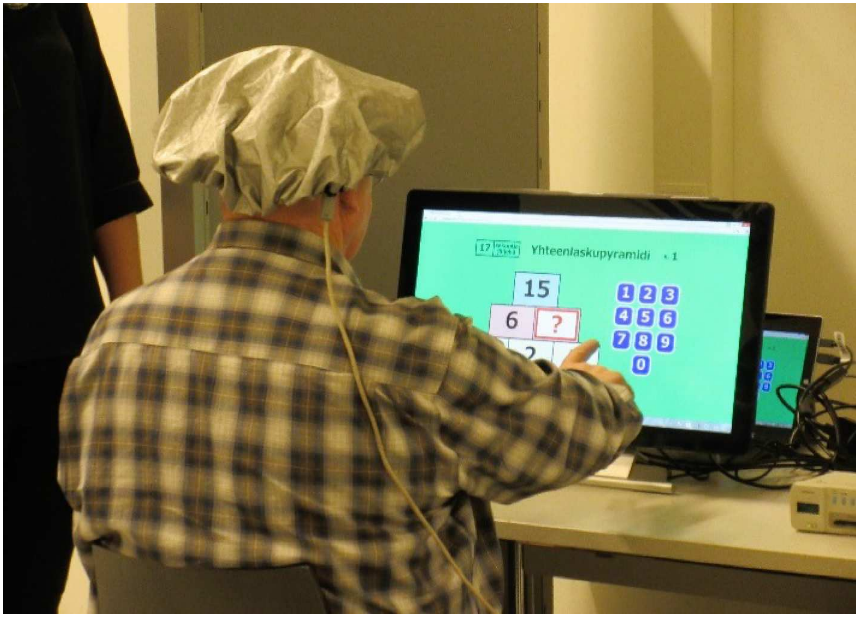


Figure 3 NIRS device being used for an elderly person form the Finnish test group.

The study employing the NIRS device as depicted in Figure 3 revealed that the elderly people within the test groups had a good and positive level of brain activity measured during the gameplay, both in the Japanese and in the Finnish test groups. This finding can lead to the conclusion that the brain trainer games may have a positive activating effect on healthy and non-diagnosed brains, which can furthermore act as one prevention method against cognitive problems, or slow down the natural ageing of the brain which can affect cognitive processes employing the working memory. The elderly people from the test groups from both countries were fluent enough with digital games and basic advice related to using the hardware devices and the game rules was sufficient. The test group in Finland had slightly more problems with the touch screen interface compared to the Japanese test group [15].

The hypothesis before the tests was that there would be more significant differences in the results from the two test groups, especially since the history of game industry and gamification is longer and more prominent in Japan compared to Finland. The quantitative data analysis did not however support the hypotheses and instead differences between the results was relatively low. The analysis of the test results was focused on the minigames which were most selected and played in both test groups, from all of the thirteen minigames. The tests revealed some cultural differences, for example the

Japanese test subjects were reluctant to choose a game that might be too difficult for them, as they did not feel comfortable losing in the game in front of other people (the test arranging personnel and the other test subjects). The test subjects from the Finnish test group preferred to try out the different minigames in order, and in the test conditions they were reluctant to communicate before the gameplay if they considered the game to be difficult or not. Some test subjects might have commented the game “not being fun”, and thus proceeding directly to the next minigame without having the desire to try out the more difficult levels of the same minigame [15].

One of the minigames displaying cultural differences between the two test groups was the “Rock, Scissors, Paper” –minigame, which has negated win condition for the game, therefore requiring the player to consciously attempt to lose in the game to provide the correct answer. The elderly people from the Japanese test group were familiar with the game, however they were not eager to play the negated win condition version of the game, because this would compel them to “lose” in this version of the game. A large portion of the test subjects from the Finnish test group were not familiar with the Rock, Scissors, Paper –game at all. Despite this they displayed eagerness to at least try the minigame out [15].

3.7. Literature review on Motivational Factors for Elderly Stroke Patients

We conducted a literature review on the motivational factors for stroke patients in the context of rehabilitation. Researchers advocate that playing games can have positive impacts and improve the quality of life of the elderly. Digital games can be used as an enhancement of mental and physical activities as well as a socialization tool for elderly. In addition, the motivational levels of the elderly stroke patients can be increased by using rehabilitation systems based on augmented reality. Based on the findings from the literature review, the researchers recommend game design guidelines for physical rehabilitation games for stroke patients [16, 17].

Researchers indicate that social functioning (e.g. social connection and contact) plays an important role in motivating patients to be engaged in their rehabilitative exercises. The relationship between patient and therapist is also an important factor. Setting a relevant personal goal is an important factor to motivate stroke patients in rehabilitation. The researchers also indicate that the rehabilitative setting and environment should be taken into consideration as an important motivational factor for patients. In stroke rehabilitation, it is important for patients to receive adequate information from the healthcare professionals (e.g. doctors, nurses, and therapists) so that they can engage in the rehabilitation process. Furthermore, physical tasks should be meaningful for patients to perform (e.g. Activities of daily living) so that they can feel engaged in these activities. The researchers also point out the importance of customization and personalization in stroke rehabilitation. There are other motivational factors for stroke patients recommended by the researchers such as positive feedback from therapists, music, and recreational activities [16, 17].

Based on the motivational factors, the researchers recommend game design principles. For social connection and functioning, multiplayer games can be designed to motivate stroke patients. A virtual therapist in digital games can be designed to promote the relationship between patients and therapists. Goal-oriented digital games can be achieved by designing effective game level design. The meaningful rehabilitative tasks can be achieved by designing real-world activities such as sports, gardening, cooking, and cycling. Game feedback, help system, and tutorials can be designed to provide effective information for stroke patients. In digital games, providing positive feedback to patients can improve their motivation in rehabilitation. The researchers also recommend other game design guidelines such as effective game music, recreational activities, and game context. These design guidelines can provide insights into designing effective digital games to motivate stroke patients [16, 17].

4. Conclusion

In this paper, we have presented the results of testing various rehabilitative games for elderly in Finland and Japan. All the tests were part of the GSH project where gamified services for promoting exercise and enhancing quality of life are researched and developed for elderly people. First, we presented the results of the Japanese usability evaluation of a prototype exergame called “The Glider”. Then, we explained the process and the results of testing SportWall game and two commercial physical activity games with Finnish elderly participants. We also presented our findings from the literature review and interviews on potential users on existing games for seniors. Next, we described the need for enhancing mental health and preventing memory disorders by proposing the “Old Photos on Map” web application where users can upload and view old photos. Furthermore, we presented the results of the usability testing of the exergames SportWall and PhysioWall that was conducted in Finland. We also suggested design guidelines for usability practitioners according to our observations from testing the Skiing Game in Japan. Then we describe the experiments conducted for the Brain Trainer Exercise game in Finland and Japan. Finally, we presented the findings of our research on motivational factors for the elderly and we recommended game design principles. We summarize the amount, age and country of the participants in each test case in Table 1.

According to all of our tests we can say that the elderly enjoy playing exergames, and we conclude that digital games can be an effective way to enhance the quality of life of the elderly. In general, the elderly gave positive feedback for the idea of using digital activity games for exercising. Especially the Skiing game was liked a lot. However, we need to take into consideration motivational, usability and safety factors when designing rehabilitative games for the elderly. In the case of “The Glider” we observed that the players where motivated while playing but motivation should also be maintained throughout the gameplay. Moreover, the elderly prefer control-free devices such as Kinect when they play. The game content and context should be simple and uncluttered. Also, the motion controls in the gameplay should be simple and jumping should be avoided for safety reasons. When talking about web applications such as “Old Photos on Map”, the elements of the user interface should be simple also large text is preferred. In addition, the elderly should receive adequate information before starting to play. Furthermore, unfamiliar activities in the game concepts should be avoided.

Table 1. Test participants in each test case

Case	**Participants**	**Age (years)**	**Country**
The Glider	12	18-23	Japan
SportWall & 2 commercial games	5	65-85	Finland
Old Photos on Map	45	30-93	Finland
SportWall & PhysioWall	19	61-83	Finland
Kinect	8	64-78	Finland
Brain Trainer game	46	> 60	Finland
Brain Trainer game	46	> 60	Japan
Skiing game	24	60-85	Japan
Skiing game	21	60-85	Finland
RecReha	10	71-78	Finland
Experiments in Singapore	35	> 65	Singapore

The arranging of the field tests is challenging. Getting suitable elderly for testing physical activity games was not easy. From arranging the field tests we have also learned that more playing and less time filling questionnaires would be more beneficial for this kind of research, suggesting for future research projects that more automated means of data gathering and analytics would bring added value to both the test arranging personnel and the test subjects.

In the future, we intend to publish the results of the following experiments. In winter 2016, we tested the Japanese RecReha game in Finland. The participants of the experiments were 10 Finnish seniors who played the game individually and in groups of 5 people. The game utilizes Kinect and the concept of the game is to try to hit virtual balls doing different kinds of movements. In spring 2016, we collaborated with Nahyang Technological University in Singapore and conducted two usability studies which include both Finnish and Singaporean game-based physical exercises. In the first study, called "Gamified solutions in healthcare and rehabilitation. Pilot study" we used five different exergames including the Finnish Skiing game and Hiking game as well as the Singaporean Chinatown game and Japanese RecReha game. The main objective of this study is to understand the usability and the user experiences of the elderly in playing five different exergames. In the second study, called "Gamified solutions in healthcare and rehabilitation – special interest in therapeutic exercise" we conducted a six week intervention study where the aim was to investigate the differences between three groups that used different exercise methods. Each group had 10 participants. The first group exercised using exergames, the second group did conventional exercise and the last group continued their life normally.

Acknowledgement

We would like to thank all of our collaborators and partners in the GSH project. This work was supported by the Finnish Funding Agency for Technology and Innovation (Tekes), City of Turku, and several industrial partners. We would also like to thank all those who participated in the game development and testing including the elderly participants and the staff from the service homes.

References

[1] R. Raitoharju, M. Luimula, A. Pyae, P. Pitkäkangas and J. Smed: Serious Games and Active Healthy Ageing: A Pre-study. in Proceedings of the 5th International Conference Well-being in the Information Society, 2014
DOI: 10.1007/978-3-319-10211-5_16

[2] United Nations, Department of International Economic and Social Affairs, Population Division. Concise report on the world population situation 2014, ST/ESA/SER.A/354. ISBN 978-92-1-151518-3., New York, 2014

[3] E. Orsega-Smith, J. Davis, K. Kelley Slavish and L. Gimbutas: Wii Fit Balance Intervention in Community Dwelling Older Adults. in Games for Health Journal, 2012
DOI: 10.1089/g4h.2012.0043

[4] W. R. Boot, M. Champion, D. P. Blakely, T. Wright, D. J. Souders and N. Charness: Video games as a means to reduce age-related cognitive decline attitudes, compliance, and effectiveness. Frontiers in Psychology, 2013
DOI: 10.3389/fpsyg.2013.00031

[5] J. A. Anguera, J. Boccanfuso, J. L. Rintoul, O. Al-Hashimi, F. Faraji, J. Janowich, E. Kong, Y. Larraburo, C. Rolle, E. Johnston and A. Gazzaley: Video game training enhances cognitive control in older adults. Nature, 2013
DOI: 10.1038/nature12486

[6] S. Kühn, T. Gleich, R. C. Lorenz, U. Lindenberger and J. Gallinat: Playing Super Mario induces structural brain plasticity gray matter changes resulting from training with a commercial video game. in Journal of Molecular Psychiatry, 2014
DOI: 10.1038/mp.2013.120

[7] A. Nakai, M. Luimula, S. Hongo and H. Vuola: Evaluating a Game Motion-Based Control by Using Kansei Engineering Knowledge. in Proceedings of the 3rd IEEE Conference on Cognitive Infocommunications, pp. 139-144, 2013
DOI: 10.1109/CogInfoCom.2013.6719229

[8] A. Nakai, A. Pyae, M. Luimula, S. Hongo, H. Vuola and J. Smed: Investigating the Effects of Motion-based Kinect Game System on the User's Cognition. in International Journal on Multimodal User Interfaces, pp. 403-411, 2015
DOI: 10.1007/s12193-015-0197-0

[9] A. Pyae, M. Luimula and J. Smed, "Investigating the Usability of Interactive Physical Activity Games for Elderly: A Pilot Study," CogInfoCom 2015 - 6th IEEE International Conference on Cognitive Infocommunications, pp. 185-194, 2015
DOI: 10.1109/CogInfoCom.2015.7390588

[10] A. Pyae, R. Raitoharju, M. Luimula, P. Pitkäkangas and J. Smed: Serious Games and Active Healthy Ageing: A Pilot Usability Testing of Existing Games. in International Journal of Networking and Virtual Organisations, 18p, 2016
DOI: http://dx.doi.org/10.1504/IJNVO.2016.075129

[11] A. Yoshii, H. Malmivirta, M. Luimula, P. Pitkäkangas and T. Nakajima: Designing a Map-Based Application and a Conversational Agent for Addressing Memory Problems. in Proceedings of the 17th International Conference on Human-Computer Interaction, 2015
DOI: 10.1007/978-3-319-21380-4_58

[12] H. Malmivirta and S. Kivelä, "Yellow Cottage and a Patch of Potato," in Art and Culture - Keys for better Brain Health. Course Material from Turku University of Applied Sciences 102, 2015, pp. 26-93

[13] T.N. Liukkonen, T. Mäkilä, H. Ahtosalo, T. Heinonen, R. Raitoharju and P. Pitkäkangas: Perceptions of the Elderly Users of Motion Tracking Exergames

[14] A. Pyae, M. Luimula, T. Saarenpää, P. Granholm and J. Smed: When the Japanese Elderly Play a Finnish Physical Exercise Game: A Usability Study. in Journal of Usability Studies, 22p (accepted)

[15] N. Katajapuu, P. Granholm, M. Hiramatsu, E. Ishihara, J. Hirayama, P. Pitkäkangas, P. Qvist and M. Luimula: Brain trainer exercise game. Field tests in Finland and Japan. in Proceedings of the International Journal of Chemistry and Chemical Engineering Systems, pp. 39-45, 2016

[16] A. Pyae, M. Luimula and J. Smed: Understanding Stroke Patients' Motivation for Motivation-Driven Rehabilitative Game Design. in Proceedings of the International Conference on Pervasive Games, pp. 99-111, 2014
DOI: 10.1007/978-3-319-19656-5_16

[17] A. Pyae, M. Luimula and J. Smed: Rehabilitative Games for Stroke Patients. EAI Endorsed Transactions on Pervasive Games, Vol. 1/4, 11 p, 2015
DOI: 104106/sg.1.4.e2

Acta
Technica
Jaurinensis

Vol. 10, No. 1, pp. 50-58, 2017
DOI: 10.14513/actatechjaur.v10.n1.414
Available online at acta.sze.hu

Prediction of Electromagnetic Fields around High Voltage Transmission Lines

G. A. Kulkarni[1], W. Z. Gandhare[2]

[1]Research Scholar, Govt. College of Engineering, Dept of Electrical Engineering, Osmanpur, 431005 Aurangabad, India
E-mail: girish227252@rediffmail.com

[2]Principal, Marathwada Mitra Mandal's Institute of Technology, Lohgaon, Pune, 411047, India
E-mail: wz_gandhare@yahoo.co.in

Abstract: The electric and magnetic fields are present around the High Voltage Transmisssion Lines (HVTL) and reported to affect health of the workers working on these hotlines. The key parameter reportedly responsible for detrimental health effects, are internal induced fields in body. Induction of internal fields in different organs is heavily dependent on the external fields. Prediction of hazardous levels of external fields before measurement or in situations difficult for direct measurement will lead to identify the restrictive situations and working conditions for hotline workers. This work propose a method to model electric and magnetic field for different climbing routes using hybrid technique, formed by combining support vector machine (SVM) and neural network (NN) and also electric field and magnetic field values are predicted using NN for increase in tower height. The result shows the performance of proposed method for prediction of electric field and magnetic field for increase in tower height.

Keywords: climbing routes, support vector machines, neural network, hybrid data mining approach

1. Introduction

Rapid expansion of power systems is causing transmission technology to drift from HV to EHV. This HVTL network is expanding with a great speed along and above the ground. Current and voltage limits are the two significant factors of high voltage transmission line [1]. Large transmission line configurations with high voltage and current levels produce large values of electric and magnetic fields stresses which influence the humans and nearby objects sited at ground surfaces. This has in turn prompted increased activity in the documentation of calculation methods to exactly forecast field strengths in isolated conducting bodies associated with lines of all voltages and design configurations [2]. Due to the extensive use of electricity in the modern domestic and industrial environment, any or all reports purporting to exhibit that the electric fields from power lines that cause or aggravate infirmity must be given serious consideration and be seriously evaluated [3].

The exposure of general public and power-line workers to high-voltage transmission lines at extremely low frequency (ELF) of 50 or 60 Hz may cause very severe health problems [4]. The human body is always vulnerable to electromagnetic radiation of varying intensity depending upon the locality.

Modeling of external fields was done using Regression analysis, SVM, NN and Hybrid Techniques [5]. In this paper a hybrid technique is used to model electric field and magnetic field present around hotline workers body while climbing through different climbing routes in a tower and also electric field and magnetic field was predicted as increase in the tower height.

2. Modeling and Prediction using Hybrid Technique

Extensive efforts carried out, to evaluate external exposure conditions and its effects reported in literature can be found out elsewhere [6, 7]. In almost all these attempts to determine the detrimental effects, electromagnetic fields are measured along the span of the line. Reducing distance between source and object, while doing live line repairs has also attracted attention of researchers to determine biological effects initiated by electromagnetic fields [6]. Prediction of fields during different positions of HVTL tower climbing routes will be helpful to avoid extreme exposure prone spots during live line maintenance.

Artificial neural networks, SVM along with classical statistical procedures such as polynomial regression are used for function approximation. These methods have succeeded in generating accurate plots, provided the input data is sufficient.

Recently, Hybrid data mining approaches have gained much popularity. A hybrid approach is built by combining two or more data mining techniques.

Reviewed literature indicates that Hybrid models can outperform standalone models and can provide better performance. The objective behind formation of Hybrid model is to extract all possible good features of individual model to obtain best possible outcomes. Generating new hybrid model by mixing individual models has also been suggested by Xu, Kryzak, and Suen (1992) [8].

Development of hybrid models by combining different NNs architectures were also reported and showed that the combinations provided improved performance compared to standalone NNs models. Hybrid models by combining statistical models (mixed regression models) have been implemented. Results from these studies suggest that hybrid models improve predictive performance when compared against the predictive performances of standalone models. The recent trend in hybrid model development is to mix classical statistical models with NNs models [9].

On modeling side, a need persists to develop a technique with conventional methods to model the external field values. There is a scope identified to implement a hybrid technique which can outperform the advantages of standalone techniques such as Neural Network and Support Vector Machines. Possibility of developing a hybrid technique using NN and SVM is exploited in this work for important issues of electromagnetic fields.

Here a new method is proposed for modeling of the electric field and magnetic field in different climbing routes of a tower and also to predict the electric field and magnetic field, if height of the tower also increases. The modeling of electric field and magnetic field is performed by using various functions such as basic fitting function, support vector machine (SVM) and neural network. Prediction of electric field and magnetic field is done for increase in tower height.

3. Methodology

In proposed method curve fitting technique is used for the modeling of electromagnetic field for various climbing routes.

Curve fitting is a process of generating a curve which best represents the characteristics of a system using the input data set. It is the basis of any analytical, comparative or growth related statistics. The objective [10] of curve fitting is to select parameter values which minimize the total error over the set of data points being considered.

The error is calculated by observing the vertical distance between the line and the point (x_{i}, y_{i}), which is given by [14]

$$E_i = |y_i - mx_i - b| \tag{1}$$

The idea behind the least squares method is to sum these vertical distances and minimize the total error. The least square error is

$$E(m,b) = \sum_{k=1}^{n} (y_k - mx_k - b)^2 \tag{2}$$

In order to maximize or minimize a function of multiple variables, we compute the partial derivatives with respect to each of the variables and set them equal to zero. Here, we compute

$$\frac{\partial E}{\partial m} = -2\sum_{k=1}^{n} x_k(y_k - mx_k - b) = 0 \tag{3}$$

$$\frac{\partial E}{\partial b} = -2\sum_{k=1}^{n} (y_k - mx_k - b) = 0 \tag{4}$$

which can be solved as a linear system of two equations for the two unknowns m and b. Since m and b are uniquely determined, these values yield the minimum error. Similarly we can proceed for any polynomial. For second degree polynomials of the form

$$y = a_0 + a_1x + a_2x^2 \tag{5}$$

the error becomes

$$E(a_0, a_1, a_2) = \sum_{k=1}^{n} (y_k - a_0 - a_1x_k - a_2x_k^2)^2 \tag{6}$$

Initially the electric field and magnetic field values for different tower height which are also different for climbing routes are taken and using this data the curve is plotted by using basic curve fitting function. The experimental dataset used to plot curve in the proposed method is shown in equation 7 & 8.

$$D_E = \begin{bmatrix} H_1E_1 \\ H_2E_2 \\ \vdots \\ H_nE_n \end{bmatrix} \tag{7}$$

$$D_M = \begin{bmatrix} H_1M_1 \\ H_2M_2 \\ \vdots \\ H_nM_n \end{bmatrix} \tag{8}$$

The above dataset is given as an input data to the basic curve fitting function and then curve is plotted for linear, quadratic, cubic, 4th degree polynomial, 5th degree polynomial, 6th degree polynomial, 7th degree polynomial, 8th degree polynomial, 9th degree polynomial and 10th degree polynomial. After plotting the curve, the best curve is taken as the experimental data curve for different climbing routes. The next step after selection of the best curve is to train SVM and neural network to predict for the electric field and magnetic field for increase in tower height.

3.1. Prediction of Electric and Magnetic Field for Increase in Tower Height

Artificial Neural Network (ANN) plays an important role while in prediction of the linear and non-linear problems in different fields of engineering [11]. Usually neural network consists of two stages namely; training stage and testing stage. In the first stage, neural network is trained by using the training dataset and in the second stage; it provides the predicted electric field and magnetic field to the corresponding height of the tower. For the training purposes of neural network back propagation algorithm is used.

SVMs are powerful machine learning techniques for classification and regression [12]. In proposed method SVM is used for prediction of the electric field and magnetic field for increase in tower height. SVM is classified into two different types binary classifier based SVM and multiclass classifier based SVM. Here more than two classes are used, so multiclass SVM classifier is used.

The hybrid technique presented is a combination of SVM and neural network. Here, the best data from SVM and neural network are taken and curve is plotted for electric field and magnetic field. For prediction, if height of the tower is given as the input; the hybrid technique gives output as corresponding electric and magnetic field.

4. Result and discussions

The implementation of proposed method was performed by using MATLAB 7.11 and the proposed method is tested for HVTL experimental data. The HVTL experimental data is taken from [13].

4.1. Prediction and modeling for HVTL data

The HVTL experiment data for electric field and magnetic field are considered for different climbing routes in a tower. Different climbing routes [13] considered in our method are shown in figure 1.

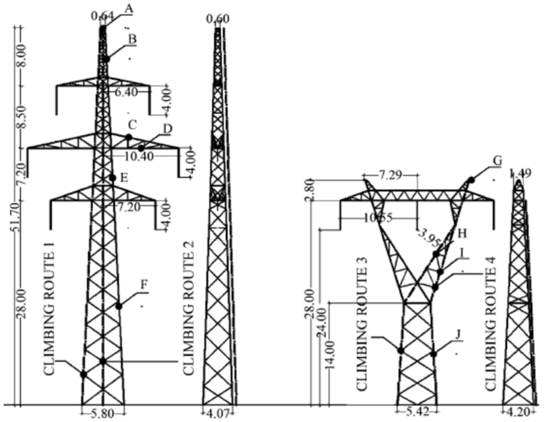


Figure 1. Different climbing routes in tower [13]

A graphical comparison of the Experimental data, fitted curves from the Regressions, the Neural Network, Support Vector Machines and the Hybrid Technique are shown in figure 2 and 3.

Modeling and prediction values for different climbing routes are shown in figure 2 and 3. The four graphs on left hand side (LHS) out of total eight shows performance of neural network, regression, SVM and hybrid technique compared with experiment data for electric field vs height from ground in climbing route 1, 2, 3 and 4. Rest of the four graphs on right hand side (RHS) represents prediction performance for both fields.

For prediction, the height of the tower is increased to 5 m from the normal height and results are analyzed. From the results it is clear that if we increase the tower height, the electric field will nearly equal to values shown in figure 2 and 3.

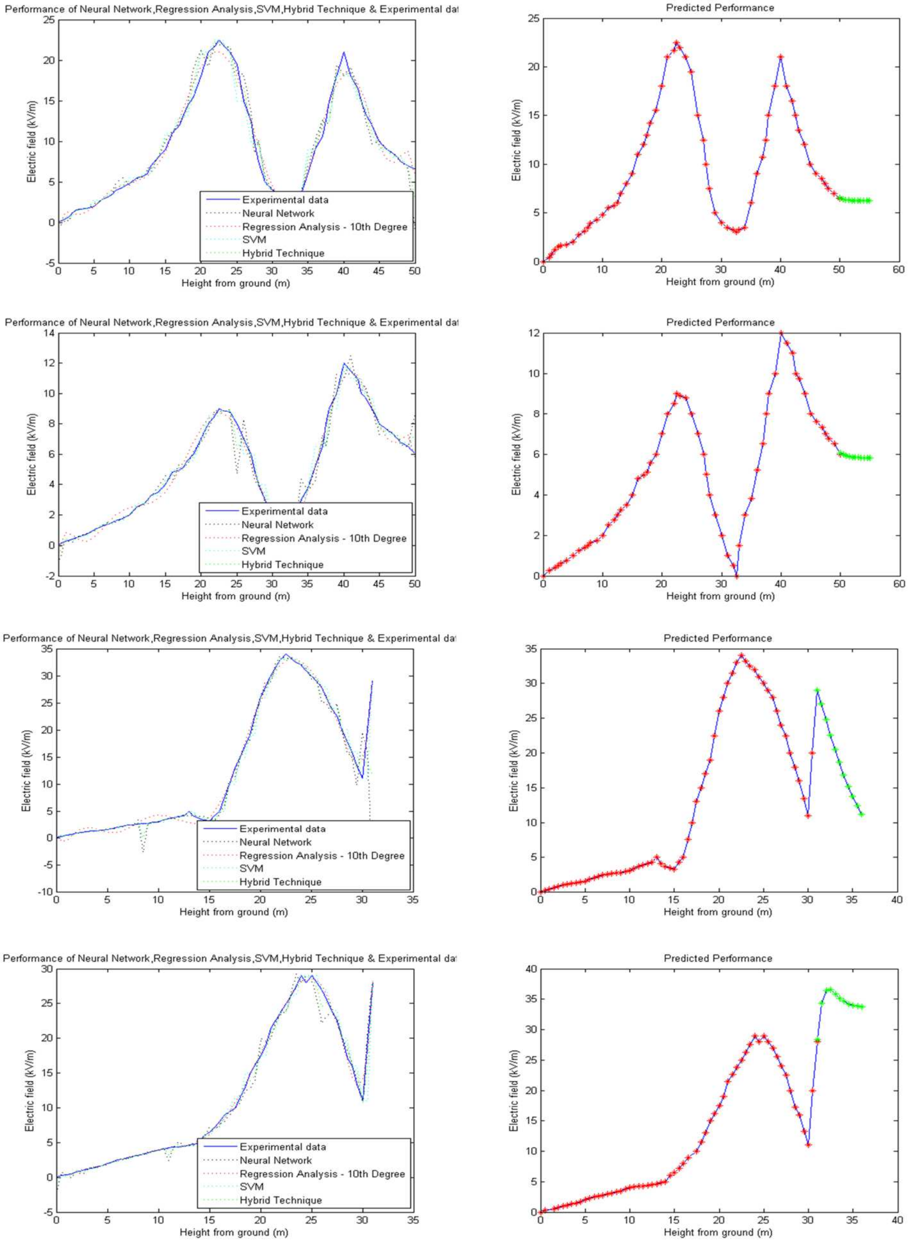


Figure 2. (LHS) Performance of neural network, regression, SVM and hybrid technique compared with experiment data for electric field vs height from ground in climbing routes and (RHS) Prediction of electric field for increase in height using proposed method for climbing routes.

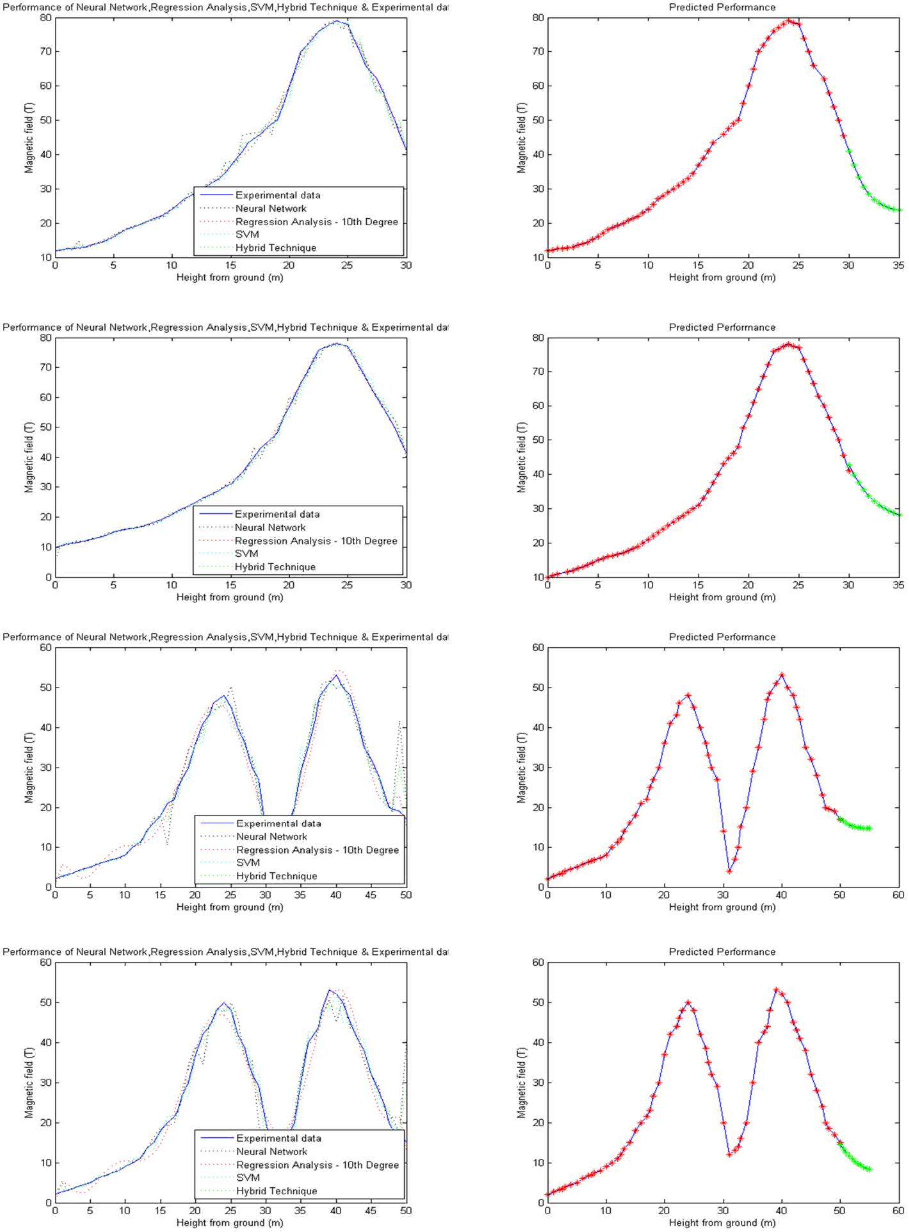


Figure 3. (LHS) Performance of neural network, regression, SVM and hybrid technique compared with experiment data for electric field vs height from ground in climbing routes and (RHS) Prediction of electric field for increase in height using proposed method for climbing routes.

5. Conclusions

Hybrid model to improve predictive performances is proposed. These gives improved results compared to the predictive performances of standalone models (NN, SVM).

The model performance of our hybrid technique was compared with SVM and NN individually. Waveforms show that the hybrid technique is better choice for prediction of EM fields around HVTL.

From the performance result it is clear that proposed method can be efficiently used to determine extensive exposure prone zones to avoid probable detrimental effects due to electromagnetic fields around hotline workers.

References

[1] Meah K, Ula S: Comparative Evaluation of HVDC and HVAV Transmission Systems. Power Engineering Society General Meeting, IEEE, pp. 1-5, June 2007.
DOI: 10.1109/PES.2007.385993

[2] Kamel NM, Gawad A: An Investigation into Magnetic Field Management under Power Transmission Lines Using Delta Configurations. The Open Environmental Engineering Journal, Vol. 2, pp. 50-67, 2009.
DOI: 10.2174/1874829500902010050

[3] Bonnell JA: Effects of Electric Fields near Power-Transmission Plant. Journal of the Royal Society of Medicine, Vol. 75, No. 12, pp. 933-941, Dec 1982.

[4] Maalej NM, Belhadj CA, Abdel-Galil TK, Habiballah IB: Visible Human Utilization to Render Induced Electric Field and Current Density Images Inside the Human. Proceedings of the IEEE, Vol. 97, No. 12, pp. 2053-2059, Dec 2009.
DOI: 10.1109/JPROC.2009.2031668

[5] Kulkarni GA, Gandhare WZ: Modeling of Electric and Magnetic Fields around 132kv Transmission Line. Acta Technica Jaurinensis, Vol. 7, No. 3, pp. 247-257, 2014.
DOI: 10.14513/actatechjaur.v7.n3.302

[6] Maalej NM, Belhadj CA: External and Internal Electromagnetic Exposures of Workers near High Voltage Power Lines. Progress in Electromagnetic Research C, Vol. 19, pp. 191-205, 2011.
DOI: 10.2528/PIERC10110601

[7] Alkoot FM, Zaeri N: Measurement of Low Frequency Electromagnetic Radiation Emitted From Overhead Power Lines in the State of Kuwait. Proceedings of the 7th WSEAS International Conference on Power Systems, Beijing, China, pp. 186-191, September 2007.

[8] Xu L, Krzyzak A: Methods of Combining Multiple Classifiers and Their Applications to Handwriting Recognition. IEEE Transactions on Systems, Man and Cybernetics, Vol. 22, No. 3, pp. 418-435, 1992.
DOI: 10.1109/21.155943

[9] Razi MA, Athappilly K: A Comparative Predictive Analysis of Neural Networks (NNs), Nonlinear Regression and Classification and Regression Tree (CART) Models. Expert Systems with Applications, Vol. 29, No. 1, pp. 65-74, 2005.
DOI: 10.1016/j.eswa.2005.01.006

[10] Balasubramanyam C, Ajay MS, Spandana KR, Shetty AB, Seetharamu KN: Curve Fitting For Coarse Data Using Artificial Neural Network. WSEAS Transaction on Mathematics, Vol. 13, pp. 406-415, 2014.

[11] Abuthakeer SS, Mohanram PV, Kumar GM: Prediction and Control of Cutting Tool Vibration in CNC Lathe With Anova and ANN. International Journal of Lean Thinking, Vol. 2, No. 1, pp. 1-23, June 2011.

[12] Woodsend K, Gondzio J: Hybrid MPI/Open MP Parallel Linear Support Vector Machine Training: Journal of Machine Learning Research, Vol. 10, pp. 1937-1953, 2009.

[13] Zemljaric B: Calculation of the Connected Magnetic and Electric Fields around an Overhead-Line Tower for an Estimation of their Influence on Maintnance Personnel. IEEE transaction on power delivery, Vol. 26, No. 1, pp. 467-474, January 2011.
DOI: 10.1109/TPWRD.2010.2064342

[14] Glenn Lahodny Jr., Curve Fitting and Parameter Estimation, Spring 2015, http://www.math.tamu.edu/~glahodny/Math442/Curve%20Fitting.pdf

Acta
Technica
Jaurinensis

Vol. 10, No. 1, pp. 59-69, 2017
DOI: 10.14513/actatechjaur.v10.n1.418
Available online at acta.sze.hu

Problems Caused by Thermal Bridges around Windows of Historic Buildings and Renovation Methods

C. Tömböly

Budapest University of Techncology and Economics,
Department of Construction Materials and Technologies
Műegyetem rkp. 3, H-1111 Budapest, Hungary
Phone: +36 1 463 3070
e-mail: tomboly.cecilia@epito.bme.hu

Abstract: Preserving the old forms and structural details in case of renovation of buildings is a primary issue. The structure of traditional wooden windows, as eyes of the building, defines the character of the facade, the street view and also influences the character of the whole townscape. With the new building energy regulations energy awareness became conspicuous, which is current requirement during the energy crisis of the XXI. century at national and international levels as well. This paper investigates all the renovation or replacement methods of historic windows should be considered.

Keywords: *energy-conscious design, historic windows, thermal performance, thermal insulation capability of windows, renovation*

1. Introduction

The building stocks of the city center of Budapest and other Hungarian cities were mainly built at the turn of the century or before, and a significant amount of them have impressive façades. Beside them, in the country there are many castles, palaces, and country houses with different architectural style. According to the data released by the Hungarian Central Statistical Office in 2003, residential buildings can be grouped according to the date of construction as the following: 10% in 1989 and 17% in 1980-89.23% in 1970-79, 15% of 1960-69, 12% in 1945-58, 12.5% in 1920-45, and 10.5% were built before 1920. It means nearly one million historic structures, and the renovation of their fenestrations is a main priority, with often preserving the inner appearance.

Instead of preserving the old forms, new characterless windows are often installed in historic buildings. The main functions of windows is to let the light in, provide air exchange and, secondly, they have to fulfill thermal and acoustic requirements. Since windows are complex structures, and they are exposed to climatic effects, their renovations have to be executed with attention. The owners have to be aware of the fact that although the reconstruction usually takes longer than the installation of a new structure, but the payback period is significantly shorter. [1]

Unfortunately, due to lack of expertise, the proper renovation of historic windows is not common, since manufacturers suggest that new products with advantageous characters are better and more modern in all cases. Old structures are often taken out and replaced by plastic windows; however, they would have been renovated and restored in many cases. The replacement causes serious damage to the character of a historic building, as well as contributes to a large amount of waste, that – from sustainability aspect – should be reused to avoid environmental contamination. [2]

2. Historic windows

The evolution of windows looks back hundreds of years. In the Baroque the windows were usually single-layered. During the classicism outward-inward opening windows with double glass layers were also applied. Since about 1855 double glazed windows with inward-inward openings appeared, but they were developed to a perfect level later in Historicism. During Secession, beside the technical knowledge, there was no limit for formal variations (Chart 1). We can conclude that these windows are style carriers, i.e. they are connected to certain architectural styles. Based on the size and forms, the age of a window can be determined with 10-20 years accuracy.

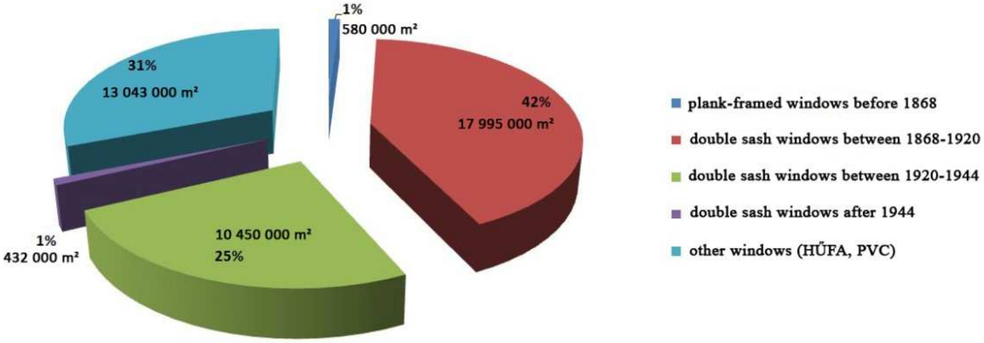


Figure 1. Windows supposed to be renovated in Hungary [3]

In Hungary the historic windows can be grouped as plank-framed and double sash windows. *(Figure 2)*. The plank-framed windows are usually found in historical buildings, and during a renovation the original forms should be saved. The double sash windows are found in a significant number of houses in Budapest and other larger cities, where the forms are the integral part of the historic cityscape. Therefore, it would be important to develop a reconstruction technology that keeps the old forms and structural design, and also meets the new technical requirements.

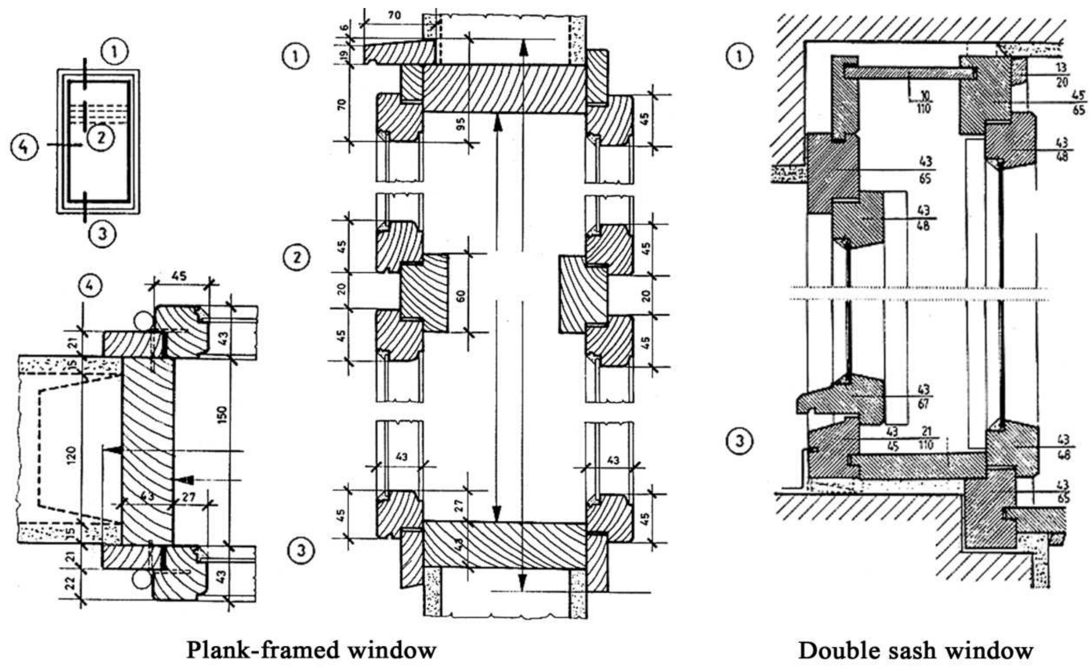


Figure 2. Details and sizes of plank-framed and double sash windows [4]

3. Calculation of thermal performance of windows and doors

During a reconstruction, decreasing the heat loss at surfaces increases the effect of thermal bridges. If the exterior walls are thermally insulated, the thermal bridges at beams above fenestrations increases. Therefore, during a renovation, instead of dealing with details, a holistic approach is necessary *[5]*.

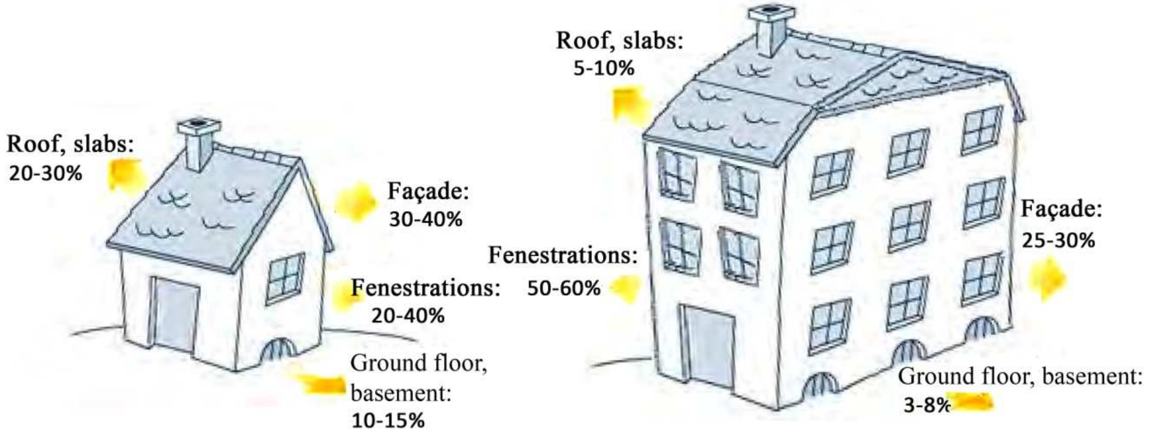


Figure 3. Heat loss at different surfaces [6,7]

Normally, doors and windows show less resistance to heat flow than other parts of the building. *[8]* During calculations the additional heat loss at thermal bridges is taken into account with the linear thermal transmittance, therefore the resulting heat transfer coefficient of a wall with windows *[9]*:

$$U_e = \frac{A_{wall}U_{layer} + \sum_j l_j \psi_j}{A} \tag{1}$$

The heat transfer coefficient of a single pane window - $U_{w.1}$ – can be calculated by the following formula *(Figure 4.)*:

$$U_{w.1} = \frac{A_G U_G + A_F U_F + l_G \psi_G}{A_G + A_F} \quad \text{[W/m}^2\text{K]} \;\; \textit{[10,11]} \tag{2}$$

where:

$U_{w.1}$ [W/m^{2}K] heat transfer coefficient of the whole window structure

A_G [m^2] surface of the glass pane (the smaller from the inner and outer panes)

U_G [W/m^{2}K] heat transfer coefficient of the glass pane

A_F [m^2] area of the frame (inner side if the window is closed, or from the outer side the larger value from the common area of the frame and the window sash)

U_F [W/m^{2}K] heat transfer coefficient of the frame

l_G [m] Perimeter of glass pane (larger one from the inner and outer perimeters)

ψ_G [W/mK] linear thermal transmittance of the glass pane

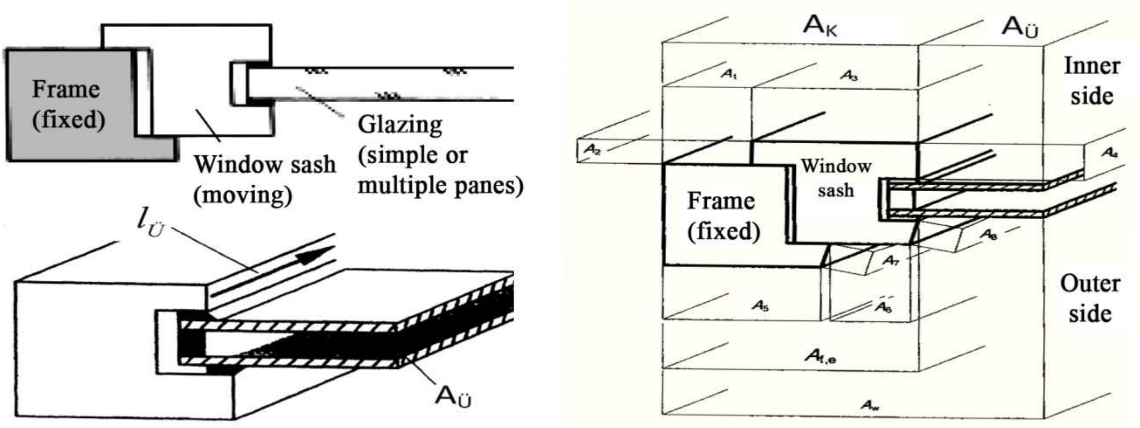


Figure 4. Different surfaces of window structures

For double paned windows - including the double sash fenestrations - the external and internal structures as well as the effect of the air gap in-between must also be considered. The heat transfer coefficient - $U_{w.2}$ – can be calculated by the following formula:

$$U_{w.2} = \frac{1}{\frac{1}{U_{w.e}} - R_e + R_{airgap} - R_i + \frac{1}{U_{w.i}}} \qquad [\mathrm{W/m^2K}] \ \textit{[12]} \qquad (3)$$

where:

$U_{w.2}$ [W/m^{2}K]	heat transfer coefficient of the whole window structure
$U_{w.e}$ [W/m^{2}K]	heat transfer coefficient of the outer window structure
$U_{w.i}$ [W/m^{2}K]	heat transfer coefficient of the inner window structure
R_e; R_i [m^{2}K/W]	thermal resistance of the internal and external glazing, not considering the other layer (these should be considered, if they are used independently)
R_{airgap}[m^{2}K/W]	thermal resistance of the airgap

In order to achieve accurate results of calculations for a certain structure, the following conditions must be considered:

- Precise geometric survey
- Determination of the correct specifications based on available literature.

4. Methods of value-added renovations of historic windows

The requirement (which is required for the heat transfer coefficient) was "k" ≤ 3.0 [W/m^{2}K] for windows and balcony doors between 1979 and 1991. A double sash window in good condition or a composite window can fulfill this requirement. This value may degrade due to the effect of filtration as function of deterioration. However, even if the filtration is decreased, i.e. the window is air-tight, better than this value cannot be reached. There was no requirements for the heat transfer coefficient of windows between 1991 and 2006. From 2006 September the TNM 7/2006 regulation provides mandatory requirements for heat transfer coefficients of windows and doors, the values are listed in following table *(Figure 5.)*:

Type of fenestration	***U_{max} (W/m^{2}K)***
Glazed fenestration on facade (wooden or PVC frame)	**1.60**
Glazed fenestration on facade (metal frame)	**2.00**
Fenestration on facade, if its area is smaller than 0.5 m^2	**2.50**
Glass wall on facade	**1.50**

Figure 5. Heat transfer coefficients of different types of windows and doors [13]

Therefore, in case of renovation of historic windows and doors, the goal is to fulfill the energy requirements of the new regulation, i.e. the heat transfer coefficient for fenestrations with wooden frames should be U_w <1.6 W / m^{2}K.

In case of existing windows these values are cannot be reached only by increasing the air tightness. More serious intervention is needed not only at joints, but also at the whole structure.

The heat transfer coefficient of a window with two glass panes without shadings (without shutters or blinds) – if it not warped and the seal is in good condition – is about 2.2 W/m^{2}K. It largely depends on the size of the glass surface relative to the wooden frame, since the wooden frame itself has very good thermal insulation capability. Thus, in case of a 50% glass/frame ratio, the U_w is 2.05 W/m^{2}K, while if the glass is 65% or more, it is 2.35 W/m^{2}K. [2] Beside this, the filtration has a significant impact as well, i.e., heat loss associated with outgoing air. In many cases, it is more important than insulating glazing, therefore it is important to seal the joints to avoid filtration*[14]*.

Possible methods of value-added renovations are the following:

4.1. Installing additional glass pane

An additional glass pane to the existing window sash can be a solution. In this case the existing window sash has to be in good condition structurally, or it should be strengthened to bear the increased burden.

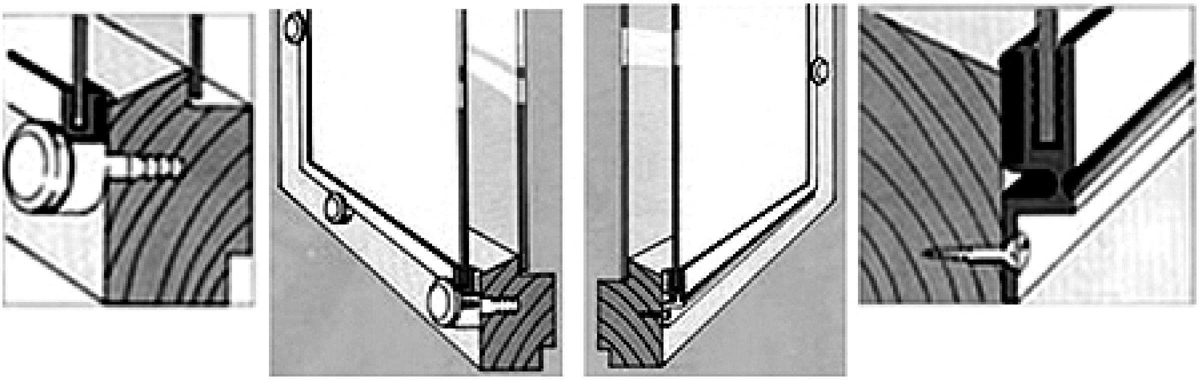


Figure 6. Additional glass pane to an existing window sash, point by point and linear (stiffer) fixing [15]

4.2. Modification of an existing window sash

Structural modification of an existing window sash can make it possible to install an insulating glazing with two or three glass layers. In this case the window sash also has to be in good condition structurally, or it should be strengthened to bear the increased burden.

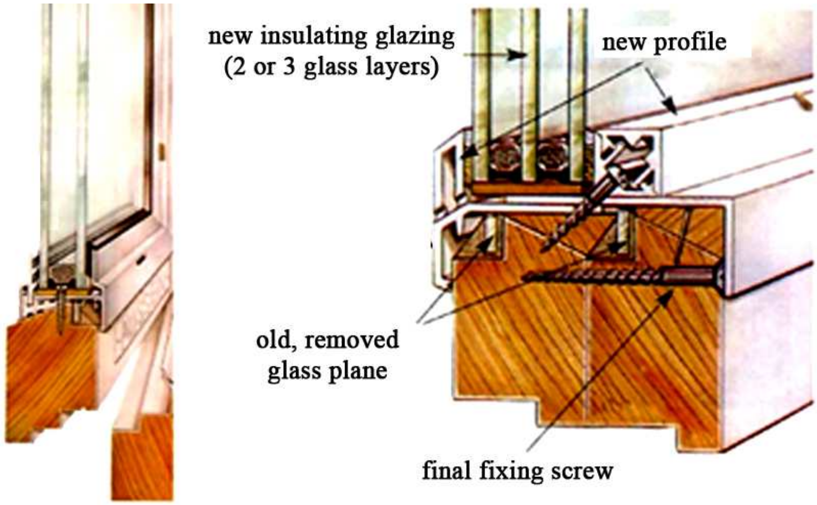


Figure 7. Modification of an existing window sash, with possible strengthening in case of installation of insulating glazing with two or three glass layers [15]

4.3. Making a new window sash with insulating glazing to the existing frame

The most common solution in case of double sash windows is to replace the glass panes of the internal window sash with insulating glazing. However, in this case, the increased weight of the glazing can be a problem. In most cases limited layer thicknesses (e.g. 3-6-3 mm) and thicker sash profile than the original one, but thinner than the ones being usual today can lead to a compromise. The suspensions should also be strengthened, and in case of bigger windows, it might be necessary to strengthen the frame as well. *(Figure 8)*.

Similar but perhaps even more combined solutions can be applied in case of plank-framed windows.

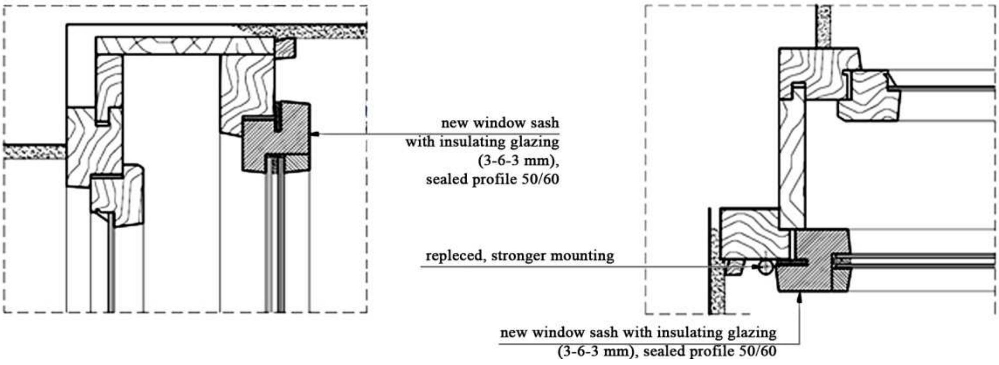
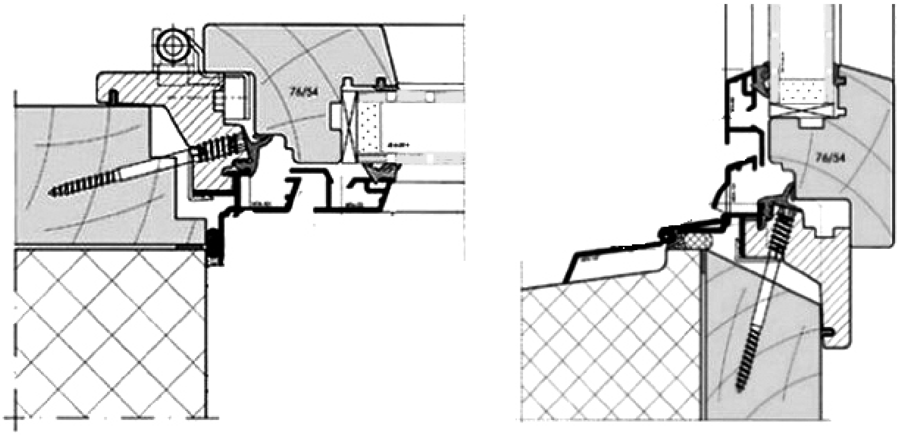

new window sash
with insulating glazing
(3-6-3 mm),
sealed profile 50/60
repleced, stronger mounting
new window sash with insulating glazing
(3-6-3 mm), sealed profile 50/60

Figure 8. Retrofitting of a double sash window with replacement of the inner window sash with insulating glazing [15]

4.4. New window sash with insulating glazing installed into the existing frame structure

In case of multi-apartment buildings, if the fenestrations are not renovated in the same time, the replacement of the internal window sashes or the internal glass panes are the only acceptable methods of renovation. This corresponds to the theoretical solutions. The outer structure may remain in its original condition. However, it is questionable, what the heat transfer coefficient of the new interior sash should be, and how it should be achieved.

Figure 9. Existing frame structure

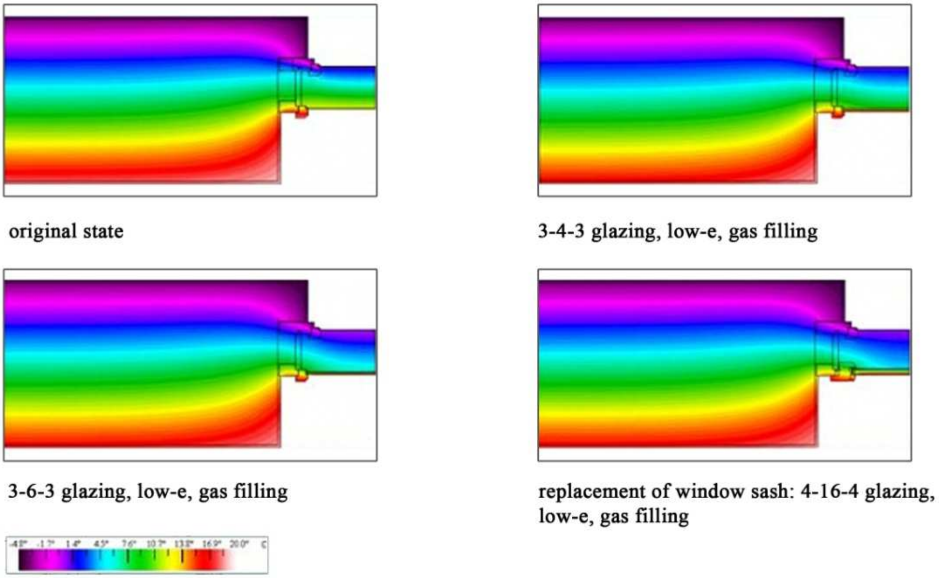


+ frame overlay and new window sash with insulating glazing [15]

4.5. Low-e glass

In case of the retention of existing windows, and the glass panes of the internal sash are replaced with a glass pane with hard low-e coating (low-radiation, invisible, with a thin layer of silver), and it is sealed at joints, we can significantly reduce radiated heat loss between the two glass layers . The low-e glass with hard coating (pyrolytic) is not vulnerable, therefore it can be used as a single layer, thus, the original sash structure does not require any changes. The new window U_w value will be around 1.8 to 1.9 W/m^{2}K, which almost meets the required 1.6 W/m^{2}K limit. This case is the least expensive solution, since the window is not modified, and the appearance does not change *[16]*.

A more efficient result can be achieved by a special glazing developed in Western Europe. It has a total thickness of 10 to 14 mm (3-4-3, 3-7-4) and it applies Histoglass D10, D14 glass panes, and the U_w is between 1.3-1.5 W/m^{2}K *(Figure 7.)* . This would result in minimal increase of the window sash, but it can be installed into existing windows sash as well, since the size of groove of the opening sash is usually 10 to 15 mm. In order to increase the thermal performance, the inner window sash can be strengthened, and with a 4-16-4 glazing the U_w can achieve 0.9-1.0 W/m^{2}K *(Figure 10.)*.


original state

3-4-3 glazing, low-e, gas filling

3-6-3 glazing, low-e, gas filling

replacement of window sash: 4-16-4 glazing, low-e, gas filling

Figure 10: Isotherms of a double sash window after renovation [17]

From structural and aesthetic point of view, these are also the best solutions. In this case the entire original window structure is retained, and the outer window sash provides an additional thermal insulation. Thus, the three layers of glass - the glazing outside and the inner double layers - already have a significantly better heat transfer coefficient than a new single pane window.

4.6. Increased thermal insulation requirements

In case of the renovation of a 2 layers window, if air-filled insulating glazing is placed into the exterior as well as into the interior window sash, the U_w <0.8 W /m^2K value can be approached, which is a requirement for the windows of passive houses. With some exceptions, the suitable air tightness can also be achieved by rubber profiles at joints. However, it can cause complete airtightness in interior spaces, therefore ventilation is inevitable. In passive houses heat recovery ventilation is successfully applied.

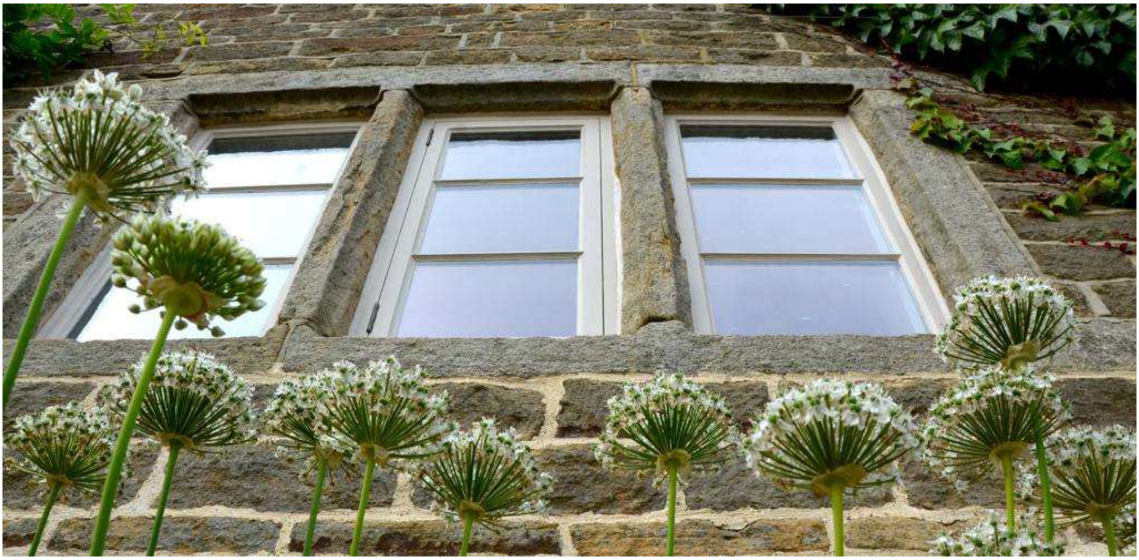


Figure 11. D10 Histoglass glazing - The Gate House - Hope Valley [18]

5. Summary

It is important to emphasize that the renovation of windows should be along with the thermal insulation of exterior walls. From the aspect of energy consumption it is important to consider the whole building, including slabs, floors as well as walls, since these are also significant cooling surfaces (*Figure 3.*). In case of historical buildings additional exterior thermal insulation often cannot be applied on facades, since it would cause the disappearance of the character of façade. (*Figure 11.*). In all cases, all the renovation or replacement methods of historic windows should be considered, since maintaining the appearance of historic buildings is a main priority.

References

[1] Trudeau P, Cambridge Historical Commission: Guidelines for Preservation and Replacement of Historic Wood Windows in Cambridge, 2009. May, www.cambridgema.gov/historic

[2] Lőrinczi Zs, The protection of historic windows, Before window to change places, www.ablakprofilok.hu

[3] Szűts L, Historical analysis of the thermal characteristics of the window structures

[4] http://www.carpenter.hu/muemlek_jellegu_nyilaszarok

[5] Vuksanovic D, Murgul V, Vatin N, Pukhkal V, Optimization of microclimate in residential buildings, Applied Mechanics and Materials Vol. 680 (2014) pp 459-466

[6] Széll M, Reconstruction Design of Buildings, Transparent facade structures of energy-efficient and sustainable renovation

[7] Matias L, Goncalves L, Costa A, Santos Ca P, Cool façades, Thermal performance assessment using infrared thermography, Key Engineering Materials Vol. 634 (2015) pp 14-21

[8] Olsen L, Radisch N, Thermal bridges in residential building in Denmark. Brno: KEA energeticka agentura s.r.o, Czech Republic, ISBN 80-902689-6-X, Available at: https://www.tc.cz/files/istec_publications/thermal-bridges.pdf (accessed 8 November 2014), 2002

[9] Blomberg, T, Heat conduction in two and three dimensions, Computer Modelling of Building, Physics Applications. Doctorate thesis, Lund University, Sweden, 1996

[10] Kusuda T, Bean J.W, Simplified methods for determining seasonal heat loss from uninsulated slab-on-grade floors. ASHRAE Transactions volume 90. part 1B, 611-632, 1984

[11] Kusuda T., Achenbach P.R., Numerical analysis of the thermal environment of occupied underground spaces with finite cover using a digital computer. ASHRAE Transactions volume 69, 439-452, 1963

[12] Rees, S.W., Zhou, Z., Thomas, H.R., Ground heat transfer: A numerical simulation of a full-scale experiment. Building and Environment 42, issue 3,1478-1488, 2007

[13] Széll M, Renovation of double-layer windows, Magyar Építéstechnika 2009/9

[14] Schittich Ch, Staib G, Balkow D, Schuler M, Sobek W, Glasbau Atlas (Birkhäuser Verlag, Basel, Boston, Berlin, 1998. ISBN 3-7643-5944-7)

[15] Tóth E, Window – People 2, Door-Window-Gate, Spektrum, Budapest, 2006

[16] Gänßmantel J, Geburtig G, Eßmann F, EnEV und Bauen im Bestand (HUSS-MEDIEN GmbH, Verlag Bauwesen, Berlin, 2006. ISBN 3-345-00873-4)

[17] Bakonyi D, Becker G, Historical windows in the light of the new requirements

[18] Histoglass Glazing Systems The Gatehouse - http://www.histoglass.co.uk/content/Project-Details/207

[19] EN ISO 10211:2007 Thermal bridges in building construction – Heat flows and surface temperatures – Detailed calculations (ISO 10211:2007). Brussels: CEN, 2007

[20] EN ISO 10456:2007 Building materials and products – Hygrothermal properties – Tabulated design values and procedures for determining declared and design thermal values. Brussels: CEN, 2007

[21] EN ISO 6946:2007 Building components and building elements – Thermal resistance and thermal transmittance – Calculation method (ISO 6946:2007). Brussels: CEN, 2007

Acta
Technica
Jaurinensis

Vol. 10, No. 1, pp. 70-83, 2017
DOI: 10.14513/actatechjaur.v10.n1.427
Available online at acta.sze.hu

Comparison of Annual Prediction Methods for Spring Flow in the Aggtelek Region

K. Mátyás[1], K. Bene[2]

[1]Széchenyi István University, Department of Transport Infrastructure
Egyetem tér 1., 9026 Győr, Hungary
e-mail: matyas.kevin@sze.hu

[2]Széchenyi István University, Department of Transport Infrastructure
Egyetem tér 1., 9026 Győr, Hungary
e-mail: benekati@sze.hu

Abstract: Karst spring flow plays an increasing important role in groundwater resources in Hungary. This paper evaluates three different estimation methods to predict mean annual spring flows in the Aggtelek region, using GIS based catchment area. Annual spring flow was predicted by two regional regression equations, by applying the Budyko equations, and by the original and modified Maucha method. Using measured spring flows, precipitation and temperature data between 1975-1992 each method was evaluated and compared for 12 spring location in the Aggtelek region. Neither method was found significantly better than the others. The Budyko curves gave a good estimation for annual spring flows, with average variance. The non-linear regression method gave the best result, with the smallest median error, and error variance.

Keywords: Aggtelek; Karst; Hydrology Statistics

1. Introduction

Understanding karst aquifers, including management, protection, and spring flow projection poses difficult challenges. There are many methods to investigate and characterize karst systems. They include: water budget, spring discharge hydrograph analyses, precipitation response analyses, deterministic (numerical modelling), lumped-parameters, and fitting statistical models [1]. Meteorological and spring flow measurements were recorded in the Aggtelek region from 1964 to 1995. This presented a unique opportunity to evaluate several karst modelling approaches to predict karst spring flow. In a previous study, Koch [2] compared the water budget method, lumped numerical models, and neural networks to predict daily, monthly, and annual spring flows. He found that the quantity of recharge changes year by year and is significantly influenced by extreme meteorological and hydrological events, large rainfalls, temperature, and vegetation changes. In an earlier study [3] the Kessler [4], Maucha [5] and Willmott [6] methods were applied to predict spring flow using recharge. The three methods gave similar prediction accuracy statistics.

Koch [2] determined that in the Aggtelek region, yearly rainfall data can be used to predict spring flow via linear regression. He used yearly rainfall data starting in July, to predict yearly spring flow starting in October. This method gave a much better estimation for yearly spring discharges (R^2=0.8) than starting both processes in January. However, these studies raised several more questions.

- can the Maucha method be improved?
- can the linear prediction equation be improved?
- why is it significantly better to use yearly rainfall starting in July, and spring flow starting in October?
- how good is the prediction if only the GIS-based catchment delineation is used?
- is the Budyko method a good alternative to predict yearly spring flows?

2. Description of the study area

The Aggtelek-mountain region is an area about 202 km^2 located in the northeastern part of Hungary, directly neighboring Slovakia, with karst cover about 70 km^2 (Figure 1.). This is a carbonate ridge of the northern Carpathian Mountains. The karst plateaus rise along an east-west direction to an elevation 400-600 m high. In this region, more than 20 springs can be found with 15 having the most significant discharge. A major portion of the karst is covered by forests containing mainly oak trees and is relatively untouched by development. It is a popular hiking destination with weather moderately cool and wet, and cold winters. The first frost occurs around October 10 and the last about April 25. The yearly rainfall is 600-650 mm with wet months June, July and August followed by two drier months, September and October. In November and December the precipitation increases slightly. Snow falls for an average 25-40 days and the snow cover remains for 60-80 days.

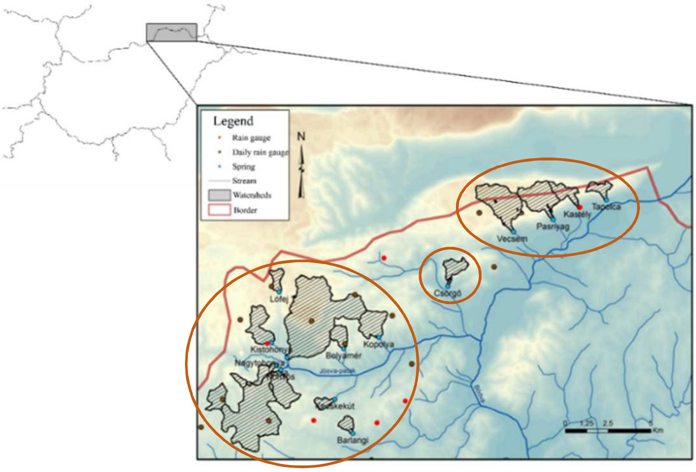


Figure 1. The Aggtelek-karst region. The springs named on the figure were monitored in a previous research (see next chapter)

2.1. Measured data

Research started in the early 1960's to study the surface and subsurface hydrology of the region. The study conducted by the Ferenc Papp Research Station included hydro-meteorological measurements, surface flow experiments and analyses, direct and indirect determination of karst infiltration, and water level measurements in karst wells. Additionally, they examined the behavior of springs by analyzing variations of their discharge and physicochemical parameters, such as electrical conductivity, and water temperature. From 1964 to 1995, 15 major springs were monitored and rainfall measurements were collected from rain gauges. The 15 monitored springs can be divided geographically into 3 distinct regions (brown ovals in Figure 1). Springs in the Jósvafő area connect into Jósva creek, except Barlangi spring joins Rét creek, and both creeks discharge into Bódva creek. Upstream on Bódva, the Csörgő and four other springs connect into Bódva stream. Although the catchments were delineated for each spring.

2.2. Geology

The primary mass of the mountain is Triassic sediment. The Lower Triassic is a clay and sandstone aquitard, overlain by clayey Middle- and Upper Triassic sediment with carbonaceous dolomite, and limestone. The impermeable Lower Triassic sediment has an overall area of 62 km^2, while the limestone-dolomite upper layers are 105 km^2. The Southwestern part of the region has a mostly impervious sediment of 35 km^2, that contains Pannonian-aged clay, sand, and gravel [7, 8]. The majority of the rainfall recharge is carried to the springs through sinkholes and the fracture systems in the karst plateau.

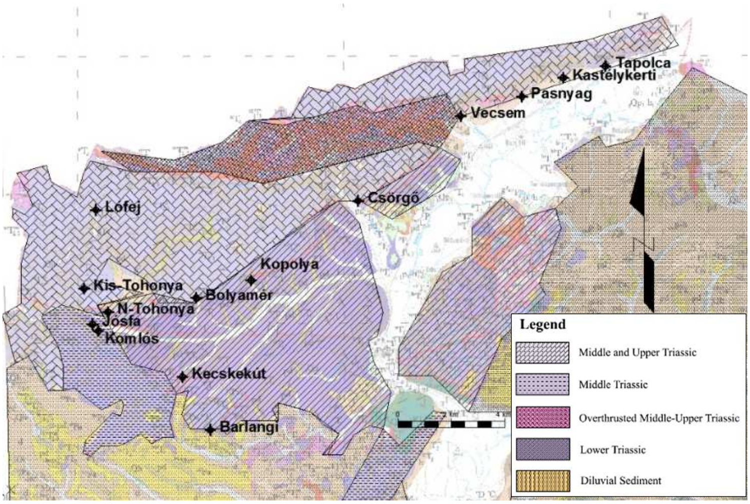


Figure 2. Geological map of the Aggtelek karst [9]

3. Time series analyses of spring flows

From the 15 springs monitored, 12 were selected for analysis. Since the springs were monitored over different lengths of time, a common time range was chosen to properly compare their hydrologic signatures. Table 1 lists the relevant data for each spring. Note that the time series used for analysis was 1974-1989 except Pasnyag. Other data include surface catchment area, mean annual precipitation (MAP), mean annual runoff (MAR), elevation of spring outflow (Meters Above Baltic Sea Level=MASL), latitude (LAT) and longitude (LON).

Table 1. Parameters for the springs

Location code	*Spring name*	*Catchment area (km^2)*	*Analyzed time series*	*MAP (mm)*	*MAR (mm)*	*MASL*	*LAT*	*LON*
1	Bolyamér	0.97	1974-1989	647	217	268	48.492	20.595
2	Csörgő	1.07	1974-1989	659	377	178	48.523	20.678
3	Kastélykerti	2.95	1974-1989	691	202	167	48.563	20.787
4	Kecskekút	0.52	1974-1989	645	228	245	48.464	20.587
5	Kis-Tohonya	2.04	1974-1989	638	278	258	48.496	20.538
6	Komlós	2.16	1974-1989	638	162	217	48.481	20.545
7	Kopolya	1.99	1975-1989	633	246	220	48.497	20.623
8	Lófej	0.67	1974-1989	645	260	428	48.522	20.545
9	Nagy-Tohonya	12.48	1974-1989	638	271	218	48.488	20.550
10	Pasnyag	5.32	1974-1985	720	718	164	48.557	20.763
11	Tapolca	1.06	1974-1989	691	206	166	48.567	20.806
12	Vecsem	5.06	1974-1989	676	164	189	48.551	20.731

3.1. Catchment area

Determination of spring flow catchment area is more complex in a karst region since many subsurface processes can influence its size. For this study, the area was based on surface geology using GIS delineation. Since most of the annual prediction methods quantify spring flow in mm/yr instead of m^3/s, accurate determination of the catchment area improves flow prediction. To assess the reliability of the GIS-measured catchment areas, the hydrologic area of each spring was calculated as well by dividing the annual spring flows (m^3/s) by the measured yearly rainfall (mm). The results are shown in Figure 3. Intuitively, catchment area should not change, however climatic factors, or long term changes in catchment characteristics can influence its size.

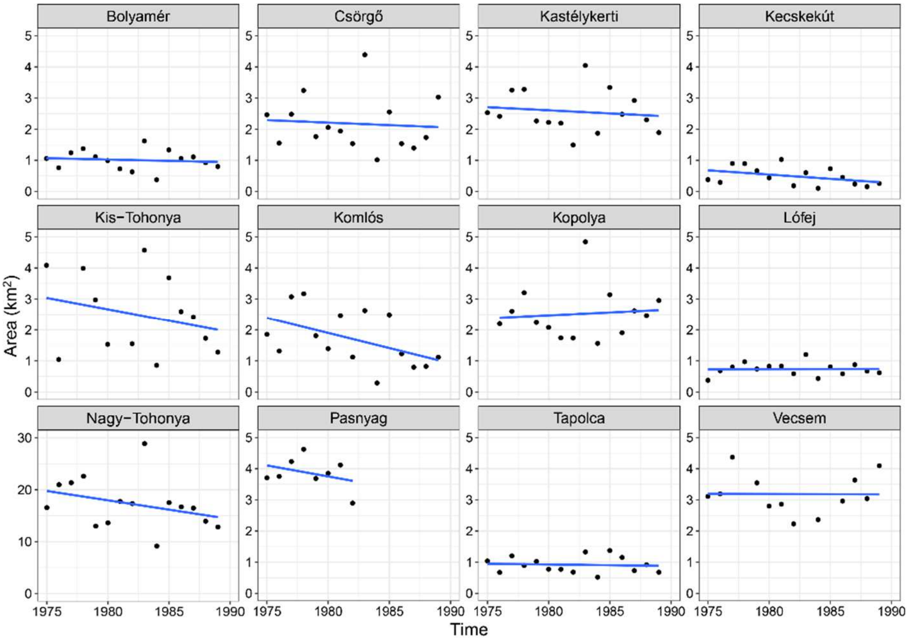


Figure 3. Calculated catchment area (A=km^2) at each spring, with trend line

Figure 3 illustrates most springs have a wide range of yearly variations. Most of the springs have a declining trend, and only a few shows (Lófej, Bolyamér) no trend. The mean calculated area was determined as well and compared to the measured values. Vecsem had a 40%, Pasnyag 31% negative difference (GIS overestimations), the rest of the springs had less than 25% over-under estimation. In this study the GIS-determined areas were used throughout.

3.2. Parde coefficient

Yearly distribution of rainfall, temperature, and spring flows were evaluated using the Parde coefficient. The Parde coefficient [10] for spring flow is calculated,

$$PCV = \frac{V_i}{\overline{V_{12}}} \ (-) \qquad (1)$$

where *PCV* is the Parde coefficient of the variables precipitation *P*, flow *Q* and potential evaporation *PET* for month *i* (-); V_i is the monthly average of variables for month *i*; $\overline{V_{12}}$ is the yearly average of variables for month *i*. The results of the non-dimensional comparisons are shown in Figure 4.

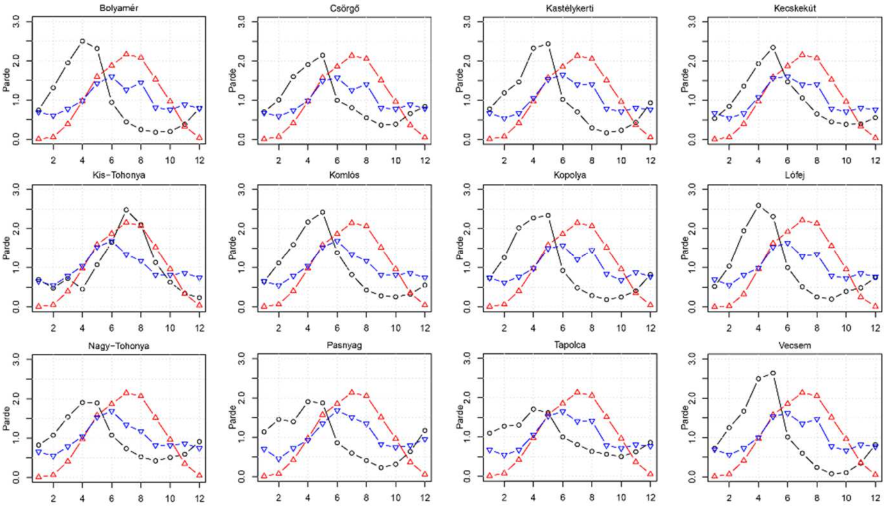


Figure 4. Computed Parde coefficients for each spring

The black lines indicate spring flow, blue lines rainfall, and red lines potential evapotranspiration. The Priestley-Taylor [11] model was used to determine potential evapotranspiration for month i,

$$\mathrm{PET}_i = \sum_{1}^{n} \alpha_{PT} \frac{\Delta}{\Delta + \gamma} \frac{R_n - G}{L} \text{ (mm/month)} \qquad (2)$$

where *n* is the number of days in month *i*; *Δ* is the slope of the saturated vapor pressure graph (kPa/°C); *γ* is the psychometric constant (kPa/°C); R_n is the available energy on the vegetation covered surface (MJ/m^{2}day); *G* is the thermal conductivity of the soil (MJ/m^{2}day, in this study *G* was assumed = 0.0); *L* is the latent heat of vaporization (MJ/m^{2}mm) and α_{PT} is the Priestley-Taylor parameter (α_{PT}= 1.26) [12, 13].

Monthly distribution of rainfall follows a similar pattern; the spring-summer months are twice as wet as the fall-winter months. Spring flows reach a maximum in April and May, and the karst system drains out in October-November. For some springs, the range between maximum and minimum flows are larger. For example, the maximum flow at the Bolyamér spring in April is twice as much as the minimum flow in October. At the Nagy-Tohonya, Pasnyag, and Tapolca springs, the distribution of flows over the year are much more uniform, indicating a hydrologically different karst system. Comparing rainfall and spring flow in the first 4-4.5 months, there is more spring flow than rainfall and the system is draining. After June, there is more rainfall than spring flow and the karst system is filling. Even when the karst system is filling, spring flows are decreasing until October, indicating a 4-month lag between rainfall and spring flow in the system. If we compare potential evaporation to rainfall at all springs, they are in phase. This reduces recharge during the wet season, and only in the winter months can significant recharge occur.

4. Annual spring flow prediction

There are several methods to predict annual spring flows. The most commonly used methods are: methods based on water budget, and two statistical methods; index methods, and regression methods. From water budget methods, the Maucha method was selected. From the index methods the Budyko method [14] was selected to predict spring flows. Using the Kessler concept (delay in the karst system), a regional regression equation was developed and an attempt to improve the prediction accuracy of spring flows in the Aggtelek region.

4.1. Maucha method

Spring discharge (*Q*) can be determined by multiplying the catchment area (*A*) with recharge (*RECH*). Recharge is the rainfall that infiltrates and recharges the groundwater. It later appears at the spring as discharge. In the Maucha method, yearly recharge (*RECH*) is calculated from the water budget components, which include raw recharge (*RECH_{raw,}*), mean precipitation coefficient (P_c), soil water storage coefficient (S_c), potential evapotranspiration coefficient (PET_c) and surface runoff coefficient (R_c).

$$\mathrm{RECH} = \mathrm{RECH}_{\mathrm{raw}} \pm \mathrm{P}_{\mathrm{c}} \pm \mathrm{S}_{\mathrm{c}} \pm \mathrm{PET}_{\mathrm{c}} - \mathrm{R}_{\mathrm{c}} \text{ (mm/yr)} \tag{3}$$

The minus sign indicates water movement out of the system. Maucha [5] determined a calculation method for each coefficient. $RECH_{raw}$ can be calculated by multiplying the monthly infiltration ratio (*IR*) with monthly rainfall, and taking the cumulative sum for the whole year.

$$\mathrm{RECH}_{\mathrm{raw}} = \sum_{i=1}^{12} \mathrm{IR} \times \mathrm{P}_{\mathrm{mon}} \tag{4}$$

In this study, a new IR was back-calculated using spring flow data over the same time period.

$$\mathrm{IR} = \frac{\mathrm{mQ}_{\mathrm{mon}}}{\mathrm{mP}_{\mathrm{mon}}} \tag{5}$$

where mQ_{mon} equals mean monthly discharge of every spring for the analyzed time period; mP_{mon} is mean monthly precipitation at every spring watershed for the analyzed time period. The new *IR* was calculated by averaging the monthly *IR* values for every spring. The Maucha and the averaged *IR* values are shown in Table 2.

Table 2. IR values for the Aggtelek region

Month	IR_{Maucha}	$IR_{averaged}$
January	0.42	0.38
February	0.70	0.68
March	0.68	0.73
April	0.52	0.68
May	0.33	0.49
June	0.16	0.24
July	0.14	0.23
August	0.11	0.16
September	0.13	0.17
October	0.15	0.17
November	0.17	0.20
December	0.30	0.34

4.2. Budyko

The Budyko method [14] was selected to describe the relationship between climate, vegetation, and the hydrologic cycle. The method expresses the dependence of actual evapotranspiration, *AET*, on energy availability (usually represented by potential evaporation) and water availability (usually represented by precipitation). Several analytical equations were developed to describe the relationship between the ratio of mean annual actual evaporation to mean annual precipitation and ratio of mean annual potential evaporation and mean annual precipitation. The equations are expressed as a function of the aridity index *Φ*, the ratio of annual potential evaporation to annual rainfall (*PET/P*). Budyko and Fu equations were selected for this study:

$$\frac{\mathrm{AET}}{\mathrm{P}} = \left[\Phi \tanh(1/\Phi)(1-\exp^{-\Phi})\right]^{0.5} \qquad [14]\ (6)$$

$$\frac{\mathrm{AET}}{\mathrm{P}} = 1+\Phi-(1+\Phi^{\omega})^{\omega} \qquad [15, 16, 17]\ (7)$$

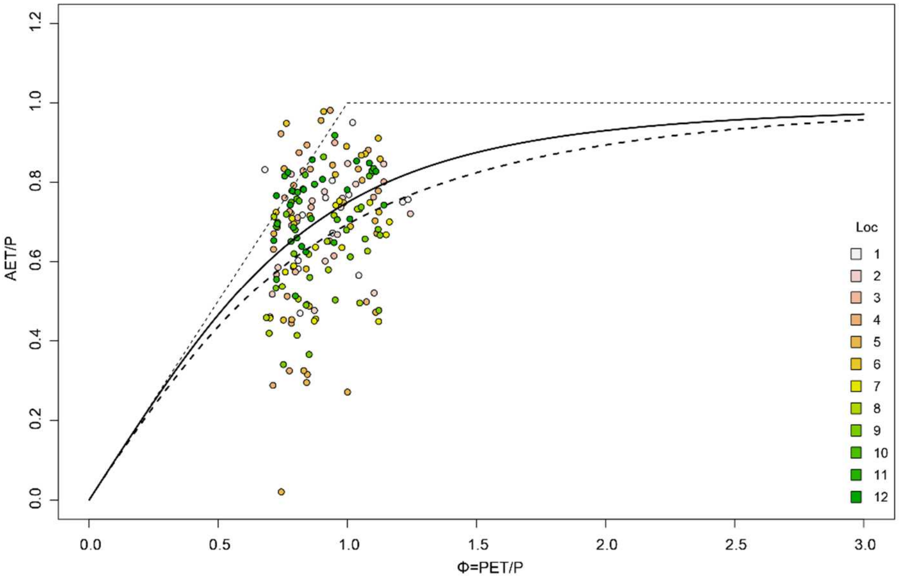


Figure 5. Budyko analysis of the time series of the twelve locations for yearly data (continuous line: Budyko; dashed line: Fu)

Where ω is a curve-fitting constant. Based on Zhang's [15] research ω = 2.4 was assumed. The Priestley-Taylor method (Eq. 2) was used to calculate potential evapotranspiration. In Figure 5 the annual aridity index and the *AET/P* ratio was calculated, and plotted for each spring. *AET* was determined with the following equation;

$$AET = P - Q \ \ (mm) \tag{8}$$

Spring locations were color coded to compare variation between spring flows for each year. The aridity index is less than 1 for most years indicating a humid climate, and the hydrological processes are energy limited. For some years the points follow the curve, but many points are above the curve indicating that potential evaporation and precipitation were in phase during the year. Using the Budyko method, spring flow can be calculated by predicting the *AET/P* ratio using one of the analytical equations, and the aridity index, then Equation 8 is rearranged to determine spring flow.

4.3. Regression equation

In a previous study by Koch [2], it was determined that better regression results can be achieved when the average yearly calculations do not start in January. A cross-correlation matrix was determined between yearly rainfall (*P*) and spring flows (*Q*), with each starting from a different month of the year. The results are shown in Figure 6. The number after the letter indicates the starting month, for example, *P2* indicates yearly calculation started in February.

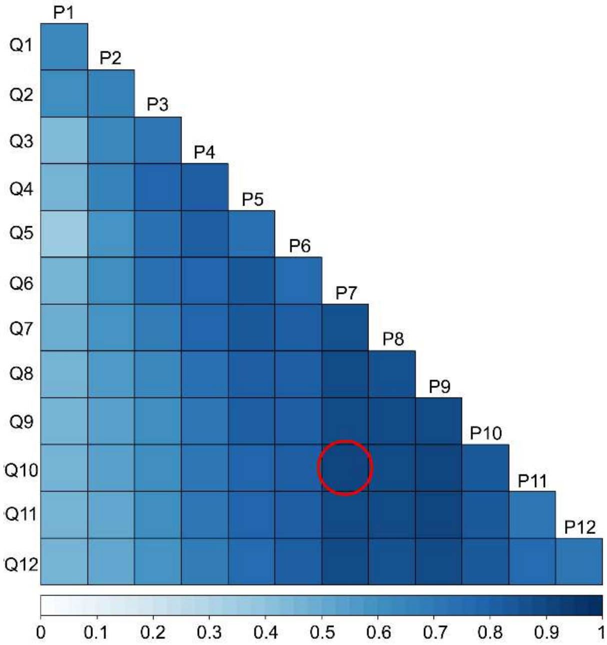


Figure 6. Cross-correlation for spring flow and precipitation

The highest correlations were found when rainfall calculations started in June, July, and August and spring flow calculations started with a 1-3 month delay. The best correlation was found when rainfall calculations started in July and spring flows in November. Using these starting points, a simple a linear regression equation was determined,

$$Q_{10} = -402 + 0.94P_7 \text{ (mm/yr)} \tag{9}$$

where Q_{10} is the annual flow starting in November. The calculated adjusted $r^2 = 0.62$. To improve prediction accuracy, several non-linear models were considered, as well as accounting for actual evaporation. The independent predictor precipitation was reduced by actual evapotranspiration (*AET*). To estimate mean annual actual evapotranspiration (*AET*), the Turc's formula [18] was applied,

$$AET = \frac{P}{\sqrt{(0.9 + \frac{P^2}{L^2}}} \text{ (mm)} \tag{10}$$

where the value of L depends on the average yearly temperature t (°C) and can be calculated as:

$$L = 300 + 25t + 0.05t^3 \tag{11}$$

The method was selected based on its reliability [19] and ease of use. For Hungary, Kovács [20] developed another estimation method to predict actual evapotranspiration. This method will be compared to the Turc method and applied in future studies.

The result of the non-linear regression:

$$\sqrt{Q_{10}} = 4 + 0.044(P_7 - AET_7)\ (\text{mm/yr}) \qquad (12)$$

with $r^2 = 0.66$. Figure 7 shows the prediction line with the 95% prediction interval and the test statistics.

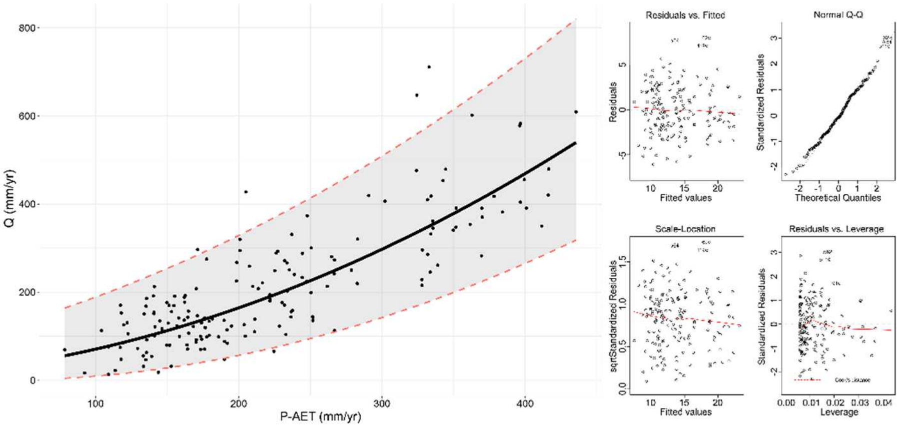


Figure 7. Results of the non-linear statistics

The test statistics are slightly better than determined in several European karst regions [21]. With no rainfall, the predicted spring flow is 16 mm/yr, resulting in constant base flow at the spring in the Aggtelek region.

5. Results and discussion

Three different modelling approaches were compared to predict annual spring flows in the Aggtelek region.

1. A water budget based method. Two variations of the Maucha method were applied: a standard variation (Maucha1) and one using a variable *IR* parameter shown in (Eq. 5) (Maucha2).
2. An index method. The Budyko curve was used to predict annual spring flows. Budyko1 was the original approach (Eq. 6) while Budyko2 used the Fu method (Eq. 7).
3. A statistical regression method. The first approach (Reg1) used a linear fit (Eq. 9) to the data, while the second (Reg2) used a nonlinear fit (Eq. 12)

To compare the three different modelling methods, the absolute normalized error (ANE) was calculated (ANE=0 for a perfect prediction),

$$\mathrm{ANE} = \left| \frac{Q_{calc} - Q_m}{Q_m} \right| \tag{13}$$

where Q_{calc} is the calculated spring flow, Q_m is the measured spring flow. The boxplot of the absolute normalized error distribution of each calculation method is shown in Figure 8.

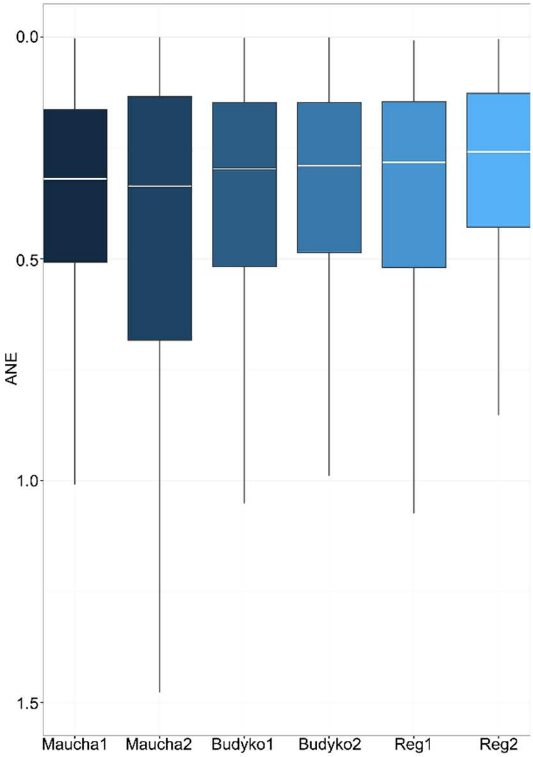


Figure 8. Comparison of 6 annual prediction methods

In Figure 8 the horizontal lines are indicating median errors, the extend of the box represent the 25-75% percentile of the errors and the upper-lower whisker extends from the hinge to the largest-smallest value no further than 1.5 * IQR from the hinge (where IQR is the inter-quartile range, the distance between the first and third quartiles). The median values in Figure 8 are all near 0.3. Median errors are lowest for the Reg2 and Maucha1 methods, while the highest is for the Maucha2 method. The Reg2 method performed best while the Maucha2 method, the worst. It appears that the original IR values are better than newly calculated IR values. The non-linear equation performed better than the linear regression. Catchment characteristics, such as elevation, and size were considered in the regression equations, but no significant correlation was found. If one considers the ease of use, the index methods performed well in comparison to the other methods. Overall the non-linear regression method gave the best prediction errors.

All methods performed similarly, since the median errors are very close. The more significant difference between the methods are in the overall spread of errors.

In this region, the aridity index is less than one, thus all catchments are energy-limited. For these types of catchments, Blöschl [22] showed similar results; in humid regions, the regression methods, and Budyko approach performs similarly. In the Aggtelek region, all methods give satisfactory annual predictions for spring flows. In future studies these predictive methods will be applied to different karst regions in Hungary.

Acknowledgement

The authors gratefully acknowledge László Maucha and Gábor Izápy for their dedicated 30 years of research and data collection in the Aggtelek karst region.

References

[1] C. J. Taylor, E. A. Greene: Field Techniques for Estimating Water Fluxes Between Surface Water and Ground Water, in Hydrogeologic Characterization and Methods Used in the Investigation of Karst Hydrology, 2008, pp. 75-111.

[2] R. Koch: Hydrological evluation of Karst-springs, Győr, 2016.

[3] K. Bene, R. Koch, G. Hajnal: Hydrological Study of the Aggtelek Karst Springs, Pollack Periodica, vol. 8, pp. 107-116, 2012.
DOI: http://dx.doi.org/10.1556/Pollack.8.2013.2.12

[4] H. Kessler: Estimation of Subsurface Water Resources in Karstic Regions, IASH II, Toronto, 1957.

[5] L. Maucha: Results and undisturbed data of karsthydrological researches on Aggtelek Hills, Vízgazdálkodási és Kutató Részvénytársaság Hidrológiai Intézete, Budapest, 1998.

[6] C. J. Willmott, C. M. Rowe, Y. Mintz: Climatology of the Terrestrial Reasonal Water Cycle, Journal of Clmyatology, vol. 5, pp. 589-606, 1985.
DOI: 10.1002/joc.3370050602

[7] L. Zámbó: The Aggtelek Karst geomorphological characterization (in Hungarian), Földrajzi Értesítő, vol. 47, no. 3, pp. 359-378, 1998.

[8] M. Veress: Factors influencing solution in karren and on covered karst, Hungarian Geographical Bulletin, vol. 59, no. 3, pp. 289-306, 2010.

[9] L. Gyalog: Explanatory book of the 1:100 000 surface geological map series of Hungary (Magyarázó Magyarország fedett földtani térképéhez 1:100000), Budapest: Magyar Állami Földtani Intézet, 2005.

[10] M. Pardé: Fleuves et rivières, Paris: Armand Colin, 1933.

[11] C. H. B. Priestley, R. J. Taylor: On the Assessment of Surface Heat Flux and Evaporation Using Large-Scale Parameters, Monthly Weather Review, vol. 100, no. 2, pp. 81-92, 1972.
DOI: http://dx.doi.org/10.1175/1520-0493(1972)100<0081:OTAOSH>2.3.CO;2

[12] T. A. McMahon, M. C. Peel, L. Lowe, R. Srikanthan, T. R. McVicar: Estimating actual, potential, reference crop and pan evaporation using standard meterological data: a pragmatic synthesis, Hydrology and Earth System Sciences, no. 17, pp. 1331-1363, 2013.

DOI: 10.5194/hess-17-1331-2013

[13] L. Y. Rao, G. Sun, C. R. Ford, J. M. Vosemodeling: Potential evapotranspiration of two forested watersheds in the southern Appalachians, American Society of Agricultural and Biological Engineers, vol. 54, no. 6, pp. 2067-2078, 2011. DOI: 10.13031/2013.40666

[14] M. I. Budyko: Climate and Life, Orlando, FL: Academic Press, 1974.

[15] L. K. Zhang, W. Hickel, W. R. Dawes , F. H. S. Chiew , A. W. Western, P. R. Briggs: A rational function approach for estimating mean annual evapotranspiration, Water Resources Research, vol. 40, p. W02502, 2004. DOI: 10.1029/2003WR002710

[16] B. P. Fu, F. B. P.: On the calculation of the evaporation from land surface [in Chinese], Scientia Atmospherica Sinica, vol. 5, no. 1, pp. 23-31, 1981.

[17] D. Yang, F. Sun, Z. Liu, Z. Cong, G. Ni, Z. Lei: Analyzing spatial and temporal variability of annual water-energy balance in nonhumid regions of China using the Budyko hypothesis, Water Resources Research, vol. 43, p. W04426, 2007. DOI: 10.1029/2006WR005224

[18] L. Turc: Estimation of irrigation water requirements, potential evapotranspiration: a simple climatic formula evolved up to date, Ann. Agron., vol. 12, pp. 13-49, 1961.

[19] J. Parajka, J. Szolgay: Grid-based mapping of long-term mean annual potential and actual evapotranspiration in Slovakia, Hydrology, Water Resources and Ecology in Headwaters, no. 248, pp. 123-129, 1998.

[20] Á. D. Kovács: Specifying lake and areal evapotranspiration rates in Hungary (in Hungarian), Budapesti Műszaki és Gazdaságtudományi Egyetem, Budapest, 2011.

[21] V. Alloca, F. Manna, P. De Vita: Estimating annual groundwater recharge coefficient for karst aquifers of the southern Apennines (Italy), Hydrology and Earth System Sciences, vol. 18, pp. 803-817, 2014. DOI: 10.5194/hess-18-803-2014

[22] G. Blöschl, M. Sivapalan, T. Wagener, A. Viglione, H. Savenije: Prediction of annual runoff in ungauged basins, in Synthesis across Processes, Places and Scales, Cambridge, Cambridge University Press, 2012, pp. 70-101. DOI: http://dx.doi.org/10.1017/CBO9781139235761.008