JOURNAL OF AGRICULTURAL INFORMATICS

AGRÁRINFORMATIKA FOLYÓIRAT

Publisher
Hungarian Association of Agricultural Informatics (HAAI)

The publishing of the Journal is supported by the European Federation for Information Technology in Agriculture, Food and the Environment (EFITA)

2017.
Vol. 8, No. 1
ISSN 2061-862X

http://journal.magisz.org

The creation of the journal has supported by the New Hungary Development

TÁMOP 4.2.3-08/1-2009-0004

"Dissemination of research result on innovative information technologies in agriculture"

The project is financed by the European Union, with the co-financing of the European Social Fund.

Nemzeti Fejlesztési Ügynökség
ÚMFT infovonal: 06 40 638 638
nfu@meh.hu • www.nfu.hu

Journal of Agricultural Informatics

Scientific Journal

Name: Agricultural Informatics

Language: English

Issues: 2-4 per year

Publisher: Hungarian Association of Agricultural Informatics (HAAI), H-4032 Debrecen, Böszörményi út. 138. Hungary

ISSN 2061-862X

PREFACE

Information technology is an everyday means that is found in all walks of life today. This is also true for almost all areas of agricultural management. The aim of this Journal is to improve scientific knowledge dissemination and innovation process in the agri-food sector. The Journal of Agricultural Informatics has been established in 2009 by the HAAI within a project of the Hungarian National Development Plan Framework. The peer-reviewed journal is operating with international editorial and advisory board supported by the EFITA (European Federation for Information Technology in Agriculture Food and the Environment).

Agricultural informatics serves not only the development of the management systems of the industry but also obtaining and publicising information on production, organisation and the market for the producer.

Technologies into network based business systems built on co-operation will ensure up-to-date production and supply in food-industry. The sector-level approach and the traceability of processed agricultural products both require the application of up-to-date information technology by actors of domestic and international markets alike.

This journal serves the publication as well as familiarization the results and findings of research, development and application in the field of agricultural informatics to a wide public. It also wishes to provide a forum to the results of the doctoral (Ph.D) theses prepared in the field of agricultural informatics. Opportunities for information technology are forever increasing, they are also becoming more and more complex and their up-to-date knowledge and utilisation mean a serious competitive advantage.

These are some of the most important reasons for bringing this journal to life. The journal "Agricultural Informatics" wishes to enhance knowledge in the field of informatics, to familiarise its readers with the advantages of using the Internet and also to set up a forum for the introduction of their application and improvement.

The editorial board of the journal consists of professionals engaged in dealing with informatics in higher education, economists and staff from agricultural research institutions, who can only hope that there will be a demand for submitting contributions to this journal and at the same time there will also be interest shown toward its publications.

Prof. Dr. Miklós Herdon

Chair of the Editorial Board

Content

Hungarian Association of Agricultural Informatics
European Federation for Information Technology in Agriculture, Food and the Environment

Journal of Agricultural Informatics. Vol. 8, No. 1
journal.magisz.org

Analysis of smallholder farmers's ICT-adoption and use through their personal information space

Mihály Csótó[1]

I N F O
Received 14 Dec. 2016
Accepted 7 Jan. 2017
Available on-line 15 Mar. 2017
Responsible Editor: M. Herdon

Keywords:
Personal information space, information sources, cluster analysis, ICT adoption, internet usage

A B S T R A C T

ICT has now apparently penetrated the agricultural production processes and farm management tasks. It is necessary to understand the characteristics of these processes in order to effectively exploit the potential inherent in infocommunication devices. The aim of the research was to explore the personal information space and the fundamental relations of information management of farmers from which important conclusions can be drawn in regard to the design, introduction and operation of information services provided to farmers and reducing the information deficit in the agricultural sector. Analysing the database comes from a questionnaire survey conducted in May and June 2015 in Hajdú-Bihar County among smallholder farmers. The article concludes that farmers have different preferences in regard to using information sources, based on which they can be divided into distinct categories, while the information space that results from their choices of these sources gives a clear picture to ICT adaptation and usage. Three distinct groups could be set up, each with their own attributes, information preferences and information activities: 'the information accumulators', 'the analytically-minded' and 'the isolated ones'.

1. Introduction

The diffusion of information and communication technologies (ICT) are taking place as we speak. These days ICT can be seen as a universal technological system which interlocks with all of the other, earlier technological systems and which has become embedded in those systems (Sasvári 2008), especially as ICT, as "general purpose" technology can be defined as a tool, a control device, organizational technology, media, and development process as well as technical practice (Molnár 2008).

ICT has now apparently penetrated the agricultural production processes and farm management tasks. However, it is necessary to understand the characteristics of these processes in order to effectively exploit the potential inherent in infocommunication devices (the application of this potential is primarily to increase efficiency). This is especially true in the case of small farms that do not and cannot maintain a separate apparatus that would carry out management tasks. It is also important to study those small-sized farms whose daily sustenance, or a significant part of it, is provided by farming (the number of such farms is in the tens of thousands in Hungary). As Szabó G. (2002) puts it, the development of information systems is a good way to reduce transaction costs, ICT is therefore crucial for smallholders, especially in terms of ex ante costs. However, since these are typically family run farms, their analysis cannot be based merely on economics as the person who runs the farm is of at least the same importance. As Szakál (1993) puts it, the family farm is a special form of joint venture, a complex entity in which business processes and satisfying the needs of the household are continuously interfering.

So it seems obvious that the use of ICT on small and family farms needs to be analysed from an information-focused perspective. As Öhlmer (1991) points out, a tool in itself is not capable of performing a miracle and the individual using the tool plays the key role: he claims that no fundamental change takes place in information processes by computerisation since that alone only

[1] Mihály Csótó
Óbuda University Digital Culture and Human Technology Knowledge Centre
csoto.mihaly@dkht.uni-obuda.hu

adds a certain level of comfort to these processes. Hill (2009) states that farmers are constantly improving their farm businesses in order to remain competitive through fine-tuning existing practices and adopting innovations, creating a unique working method where (Hill cites Vergot, Israel & Mayo 2005 and Solano, Leon, Perez & Herrero 2003) farmers (as individuals) "…*have their favoured information sources, which they use depending on the specific information being sought*".

The information-centred approach leads us to the person's (in our case: the farmer's) information culture which is (as Z. Karvalics (2012) cites Gendina (2009) "*one of the components of a person's general culture; sum total of information outlook and a system of knowledge and skills providing goals aimed at independent activity in optimal satisfaction of information needs on the basis of both: traditional and new information technologies. It is the most important factor of successful professional activity as well as social safety in information society*." Z. Karvalics also emphasis the importance of "*the personal information space*", an umbrella term, which means the continuously developing and expanding "cloud" of individually-selected content, personalized information services and advanced information management tools.

The personal information space appears in many earlier research work relating to ICT adoption in agriculture. As Harkin (2006) stated, telematics as a medium should be examined in relation to competing media and its strengths and advantages over the conventional methods of information dissemination exploited. Doye, Jolly, Hornbaker, Cross, King, Lazarus, Yeboah & Rister (2000) concluded that farmers differ in the ways they use management information, from being information "hogs" requiring great amounts of detailed information, to being "seat of the pants" decision makers where experience and intuition are all that is used in decisions. Berman (2006) stated that the efficiency of different information transfer methods depend on the ability of the end-user farmer, his or her practices in terms of problem identification and analysis, information gathering, critical thinking and evaluating outputs. Alvarez and Nuthall (2006) concluded that farmers' software adoption behaviour results from a complex pattern of interrelationships involving structural factors, such as farm and farmer characteristics as well as 'soft' variables, such as goals and practices, which 'mediate' the effects of the first ones.

Parallel with the research on the characteristics of farmers, modelling the agricultural information flow has also been at the focus of many related projects and research in the last 20 years or so (e.g. Szabó (2000) cites Kozári (1994), Sörensen, Fountas, Nash, Pesonen, Bochtis, Pedersen, Basso & Blackmore (2010), Řezník, Lukas, Charvát, Charvát Jr., Horáková & Kepka (2016)). Based on all the above information one can conclude that with the examination of the information environment or the preferred sources of farm-related information some important insights about the role of ICT in the personal information space of farmers can be gained.

2. Hypothesis, methods and material

The first objective of the research was to explore the attitude farmers have to generally used ICT innovations (computers, internet and smart phones) in Hungary. I start from the premise that Hungarian farmers are no different to the general Hungarian population in regard to their acceptance and willingness to start using these so-called general purpose technologies that can be utilised in several areas of life. The second objective, closely linked to the first one, is the examination of the information environment of farmers. The exact role of ICT means can only be fully assessed if we are familiar with the information processes in farms and the sources of information available to farmers. New technologies and solutions must be integrated into the already existing processes, thus it is crucial to know what sources are preferred by a farmer in attaining information concerning farming, and it must also be explored if distinct groups with clearly delineable attributes can be identified within the farming society based on preferred sources. Important conclusions can be drawn in regard to the design, introduction and operation of information services provided to farmers by answering the above hypotheses and by examining the spreading of innovation as well as the fundamental relations of information management.

The main hypothesis of the research was that based on their preferred sources of information farmers can be divided into distinct groups, each with specific attributes, and these groups take a different approach to the use of information technology.

I analysed a database based on a questionnaire survey conducted in May and June 2015 in Hajdú-Bihar County with the cooperation of the county directorate of the Hungarian Chamber of Agriculture (NAK). The delivery and filling in of the questionnaires by the farmers was assisted by the experts of NAK's agriculture extension network ('Falugazdász' in Hungarian). According to my previous knowledge, each extension worker has an approximately similar number of clients, so every officer distributed the same number of questionnaires and they were instructed to have the clients arriving at their next consultation fill them in (and if there are not enough clients, then have the rest filled in during the next consultation session). Hence, the surveyed population was the circle of farmers registered in Hajdú-Bihar County, and the method used a quota-based sampling combined with accidental factors.

The questionnaire had 45 questions seeking to cover all the factors considered as relevant by literature. The first section contained questions about ICT tools and internet use (also asking about the functions used and the frequency of their use in the case of mobile phones). The second section was aimed at examining the attributes of internet and computer users (the beginning of the use of the technology, accumulated experience, the evaluation of their own IT skills, the extent of support) as well as the form and frequency of use, also focusing on various agricultural software programmes and agriculture-related applications. In the case of the latter I devoted special attention to communication, information and transaction services. In this same section those who do not use the internet were asked why they opted for non-use. The third section started with questions about the sources of information necessary for farming, followed by questions in regard to the various factors impacting innovation, i.e. questions about one's social network and approach to innovation, the reliability of online content, and the perceived usefulness, the ease of usage, observability and compatibility. The last two sections of the questionnaire were devoted to the given farmer's socio-demographic and farming-related attributes and asked for the description of the farm he or she owned.

Out of the 200 questionnaires that were handed over, a total of 148 were suitable to be evaluated. I recorded the information included in the questionnaire electronically, and ran a consistency check. I then converted the records into the SPSS statistical programme, where I completed the required data cleaning tasks along with the filtering out missing/contradictory data, while altering the existing variables (necessary for logical and/or distributional reasons) into ones that can be better used in the analysis. I used a special case of factor analysis, called main component analysis, to produce factors used for the necessary data reduction, dimension reduction. Based on the results I gained during the principal component analysis of preferred media, I divided the farmers into groups with the help of cluster analysis using hierarchical clustering, which is aimed at gradually decreasing the number of groups by merging at every stage of the process those two groups that are in the closest proximity to each other and show the greatest similarity. I applied the squared Euclidean distance to determine the distance between the objects, and I chose Ward's method aimed at minimising the total within-cluster variance.

3. Results

3.1. ICT-ownership and usage among farmers

Fifty-nine percent of respondents have a desktop in their homes, while for 44 percent of them a notebook or a laptop is a more accessible solution, and tablets are used by 10 percent. A total of 80 percent of those asked have access to some kind of computer in their homes, and the proportion of those with internet access is the same. The majority subscribed to wired broadband internet (this connection is found in the homes of two-thirds (68%) of those asked).

Almost all (95%) of the farmers included in the sample have a mobile phone, and the majority use smartphones rather than traditional mobiles (49% of all farmers, of which 4% also have a traditional mobile). This roughly accords with the nationwide data for the Hungarian population. 64% of mobile-

owners have a subscription, while one third (31%) use the pre-paid scheme of telephone use, and a small group (5%) has both. The data listed here significantly overlap with KSH (Hungarian Central Statistical Office) data (households with desktop computers: 53%; households with laptops: 45%; households with internet access: 73%; households with mobile phones: 95%), which indicates that the same diffusion process is taking place among farmers as in the rest of the Hungarian population in regard to general-purpose ICT use. As regards internet use, half of the respondents use the internet on a daily basis, one fifth use it several times a week, and some ten percent use it less frequently than that.

In regard to directly applied agricultural solutions it can be stated that the use of various software programmes supporting the management of farming is also clearly present in a certain group of farmers: half of internet-user farmers use farming log-book software, while a quarter of them keep some kind of electronic register. The wide use of the farming log-book (FL) is somewhat overshadowed by the fact that keeping a log-book is mandatory to apply for certain forms of financial support. However, since these support schemes are not tied to the use of an online FL, it is primarily farmers with a basic openness to ICT who tend to choose this software. The support of e-management and decision-support systems are not yet widespread. To sum up, some form of agricultural software is used by 60% of computer-user farmers, and 46% of the whole sample (out of this 28%, 14% and 4% of the respondents use one, two or three applications, respectively).

3.2. Sources of agricultural information and the personal information space of farmers

The farmers were asked to evaluate the sources of information they can potentially access, and rate their importance in farming management. Figure 1 shows the distribution of answers provided; the numbers in brackets after the sources are the average values for the given source. The role of NAK extension workers (Falugazdász) (4.55) as sources of information outweighed all other alternatives, although it might partly result from the sampling method, since the questionnaires were filled in with the help of the extension workers, and even if some of the farmers who went to the rural consultancy office for the questionnaire research are not regular visitors to the office, it is likely that it was those farmers who ranked the role of extension workers in their personal information network in a prominent place who were included in the sample. It could be observed in earlier research (Herdon & Csótó 2009, Csótó 2013) too that extension workers are actually well equipped to provide personalised information to farmers; moreover they render assistance in transaction services and are able to synthesise the benefits of other sources.

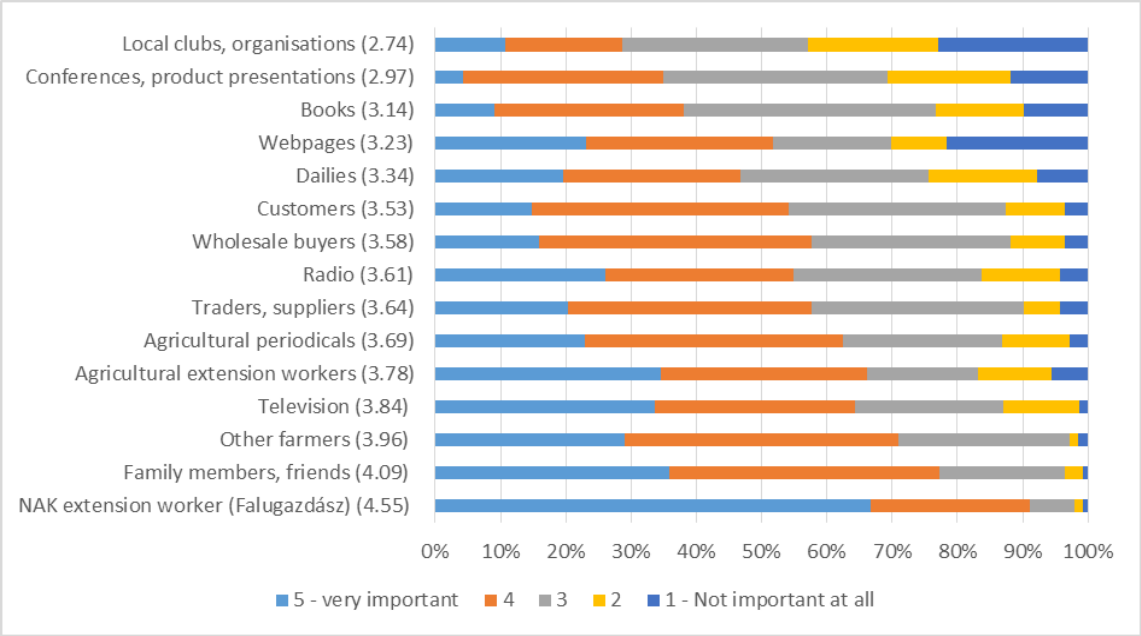

Figure 1. Assessing the importance of various sources of professional information

Television (as mass media, from where general information can be gained about important issues first) is still in the top five choices among preferred and important sources of information in addition to other personal sources of information. Agricultural periodicals also occupy a prestigious place in this regard, while books and other sources providing knowledge transfer in groups are lower in the ranking. Websites lagging behind came as a surprise, however, it must be noted that this question was also answered by non-internet-users, which negatively impacts the average of this source. After filtering out non-internet-user farmers the average for the internet in this question rises to 3.73, i.e. almost comes on a par with that of agricultural extension workers and specialist periodicals.

As could be seen in the introduction, several authors claim that there is a difference between the decision-making techniques of farmers, and this is reflected in their media use. In order to confirm this, an explorative factor analysis was conducted to decrease the dimensions of the listed sources (the Internet and the Falugazdász were excluded) and then divide the farmers into groups based on their preferred media, or preferred categories of media.

Using the Kaiser-Meyer-Olkin (KMO) statistics to discover if data are likely to factor well, the result was 0.789 which is a quite good (and well above the minimally required 0.5 or 0.6 value), and the three factors together accounted for 63% of the total variance. Table 1 shows the results of the rotated component matrix, where loadings of less than 0.4 are not presented in order to make interpretation easier.

Table 1. Rotated Component Matrix

Information sources	1st component	2nd component	3rd componentt
Television		,837	
Radio		,876	
Agricultural periodicals		,508	,460
Dailies		,730	
Other farmers	,717		
Family members, friends	,471	,495	
Books			,811
Local clubs, organisations			,683
Agricultural extension workers	,447		,495
Traders, suppliers	,808		
Conferences, product presentations	,428		,566
Wholesale buyers	,865		
Customers	,813		

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
Rotation converged in 4 iterations.

The analysis revealed the existence of three factors representing clearly distinguishable communication activities:

- Component 1: ‘Personal professional sources’, in which the personal, face-to-face dialogue plays the central role, primarily conducted with professional players (other farmers, traders, suppliers) and to a lesser extent with family members, agricultural consultants and perhaps in the framework of professional events.
- Component 2: ‘General sources’, with the key role played by traditional mass media (TV, radio, dailies (and agricultural periodicals to a lesser extent) and everyday communication with family and friends relations.

- Component 3: 'Analytical sources', featuring agricultural mass media (agricultural periodicals and books) and group activities (local clubs, conferences, product presentations).

When the different factor scores are represented according to the age groups of the farmers (Figure 2), a turning point can be observed at 40-50 years, since while the third factor dominates in younger age groups (i.e. a kind of analytical approach, which correlates with the higher level of education in the case of younger farmers), it has a negative factor score in older age groups. The middle age groups tend to avoid the middle component, i.e. general sources. The role of personal professional relations steadily increases up to age 60 (as these relationships become more extensive and trusted with time), but for pensioner age farmers the dominance of general sources can be seen.

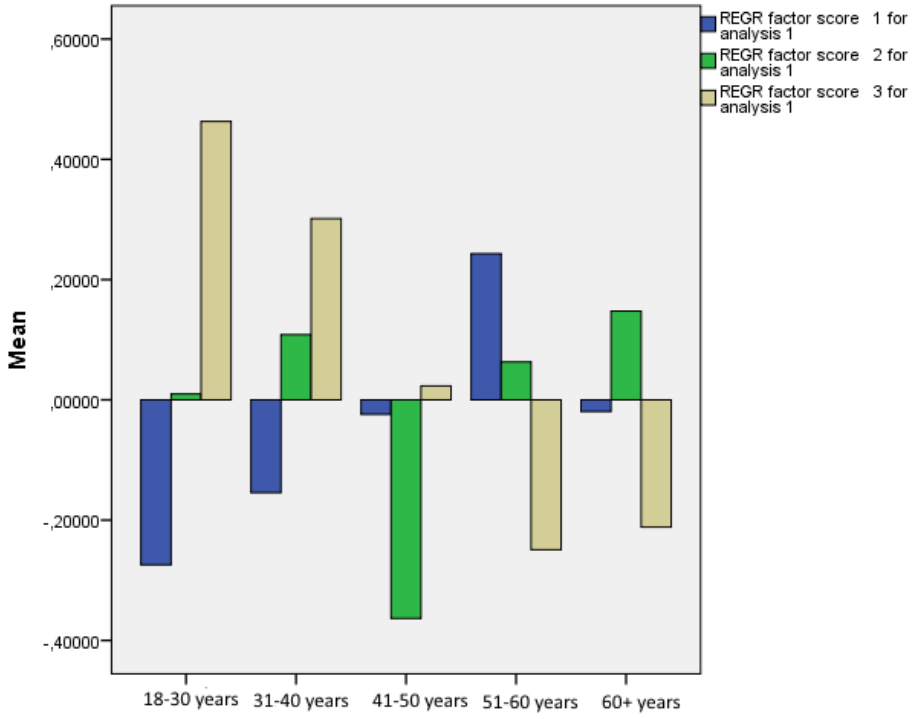

Figure 2. Distribution of media use factor scores according to age

A cluster analysis was conducted with the factor scores, and based on the result three groups can be created, the first two of which comprise ICT-active, while the third one non-ICT-active famers.

- Group 1 ('the information accumulators', 26% of the respondents): in most of the cases they are average ICT technology users; almost half of them have agricultural qualifications. Both the general sources and the analytical information factors are significant here, as the members of this group gather information from various sources.
- Group 2 ('the analytically-minded', 38% of the respondents): the most active group with a marked proportion of the middle-aged (41-50) and a significantly higher number of those with agricultural qualifications (¾ of the group has such qualifications). We can see the dominance of the analytical information factor in this group, while the use of general sources is not at all characteristic.
- Group 3 ('the isolated ones', 36% of the respondents): a significantly smaller part of this group went to computer courses; the proportion of the older age groups is some 15-20 percent higher. Only a small proportion (10%) in this group have degrees in tertiary education, and less than half of them have agricultural qualifications (45%). One fifth do not regularly discuss matters relating to their farms with anyone and almost exclusively use general sources of information, but not to a great extent.

It is important to note that there is no significant difference between the groups in regard to whether the farmers in it perform their activities full-time, i.e. the fact whether agriculture is a main source of income or not does not have an influence on their management methods and information

management (Table 2). Group 2 is a kind of 'farmer elite'. A high number of them have agricultural degrees, and, thus, a more analytical way of thinking, and they own larger and more successful farms; these factors are clearly interrelated. The 'isolated ones' are significantly older (and farming during retirement). The 'information accumulators' are a somewhat younger group of farmers, who usually have smaller farms than the other two groups.

Table 2. Demographic and farm characteristics in the identified farmer groups

	Group 1: Information accumulators	Group 2: Analytically minded	Group 3: Isolated ones
Age			
Below 40	43%	34%	20%
Between 40-60	39%	53%	49%
Over 60	18%	13%	31%
Sex			
Male	61%	72%	71%
Female	39%	28%	29%
Education (%)			
Primary	9%	4%	12%
VET	24%	13%	25%
Secondary	27%	34%	53%
Tertiary	40%	49%	10%
Agriculture qualifications (any level)			
Yes	55%	74%	45%
No	45%	26%	55%
Business orientation			
Subsistence farming	21%	9%	20%
Selling the majority of products	79%	91%	80%
Employment status			
Farming is the main activity	43%	53%	43%
Farming is a part-time activity	39%	34%	26%
Farming during retirement	18%	13%	31%
Size of farm area			
0-5 hectares	46%	13%	35%
5-20 hectares	30%	38%	33%
20-100 hectares	15%	32%	22%
100 hectares or more	9%	17%	10%

The 'analytically minded' group is the most efficient in integrating ICT solutions in their management practices, as will (Table 3). In regard to internet use, almost all the members (91%) of group 2 are internet-users, while this percentage is approximately the same for the other two groups (79% for group 1 and 71% for group 3, the latter being almost 10% lower than the average); the difference between the groups is significant. Group 3, the one less open to ICT, is significantly lagging behind in regard to mobile phone use (the percentage of mobile phone use is 61% and 66% for groups 1 and 2, respectively, while it is only 37%- for group 3).

A clearly visible and significant difference can be seen in the area of agricultural software use: groups 1 and 2, which resembled each other in many other respects, are clearly dissimilar in this regard. While two thirds in group 2 use agricultural software, the proportions are reversed in groups 1 and 3, which is reflected by the number of accessible computers, where the distribution is similar in the latter two groups: while almost all the farmers in group 2 have a computer at home, 30% do not have one in both of the other two groups. The differences also can be seen through the intensity and frequency of internet use.

The majority of agriculture-related internet activities are part of the daily or weekly routine of the 'analytically minded' farmers (averages below the value of 3 (at least monthly) and close to the value 2 (weekly) indicating this), while the 'isolated ones' perform these activities only monthly or less often (if at all), while the 'information accumulators' are somewhere in the middle of the two other groups.

Table 3. The usage of different ICT tools and services among the identified farmer groups

	Group 1: Information accumulators	Group 2: Analytically minded	Group 3: Isolated ones
Internet (%)			
Use	79%	91%	71%
Do not use	21%	9%	29%
Smart phone (%)			
Use	61%	66%	37%
Do not use	39%	34%	63%
Agriculture software (%)			
Use	33%	66%	37%
Do not use	67%	34%	63%
The frequency of different internet activities (mean, on a scale of 1 (daily) - 5 (never))			
Visiting agricultural forums, subscribing to newsletters	2,65	2,09	3,14
Looking for agricultural news	2,69	2,04	3,14
Looking for information on agricultural goods and services	2,62	2,3	3,66
Looking for information from government	3,04	2,11	3,38
Looking for information on prices	2,75	2,4	3,48
Looking for information before bigger investments	3,2	2,83	3,86
Online banking	3,08	2,65	3,28
Buying on the internet	3,38	3,6	3,97
Selling on the internet	3,81	3,83	4,38

In order to gain a deeper understanding of the groups, some latent variables regarding innovativeness and ICT-adoption were also examined. The variables (Table 4) were fine-tuned to this study using the earlier works of Aubert et al. (2012) and LaRose et al. (2012). The factor scores of each variable between the different groups are shown in Figure 3.

Table 4. The construction of the latent variables (respondents had to evaluate the statements on a scale of 5 whether they find them appropriate for themselves (5) or not (1))

Latent variables
Openness to new things (Cronbach's Alpha: 0,929)
I have always been curious about how to operate new things and innovations
I like to experiment with new solutions
If there is an opportunity, I try to grab it
I seek the company of those who are always trying out something new
I regularly look for new products and solutions
ICT self-efficacy (Cronbach's Alpha: 0,877)
Using the internet is not particularly complicated for me
I have a basic understanding how to cope with the internet

Browsing internet content and using applications that are relevant to me is not difficult
Managing mobile phones is difficult for me (reverse coding)
It is hard to learn the use of computers and cell phones (reverse coding)
ICT observability (Cronbach's Alpha: 0,773)
Using the internet helps a lot for those living around me
I heard/read good things about the internet in the newspapers, and on TV
Many people use the internet in my environment
I heard nothing but good things about the internet from my family and friends
Perceived usefulness (Cronbach's Alpha: 0,871)
Using the internet can save money
Farm-related tasks are easier and quicker with the help of the internet.
Getting information that is necessary for farming (e.g. prices, weather) is easier by using the internet
Using the internet can save time
ICT compatibility (Cronbach's Alpha: 0,787)
The computers can help me to take care of the management of the farm in the way I used to
Using the internet does not fit with the way I am doing business (reverse coding)
I think the use of computers in agriculture is straightforward

To sum up the findings, farmers have different preferences in regard to using sources, based on which they can be divided into distinct categories, while the information space that results from their choices of these sources gives a clear picture to ICT adaptation and usage. During the main component analysis and cluster analysis that were conducted on farmers' preferences of sources of information three distinct groups could be set up, each with their own attributes, information preferences and information activities: 'the information accumulators', 'the analytically-minded' and 'the isolated ones'. Those in the first group are active users of sources of information, characterised by an openness to ICT, although they do not integrate ICT into their farm management activities. The second and third groups are each other's opposites.

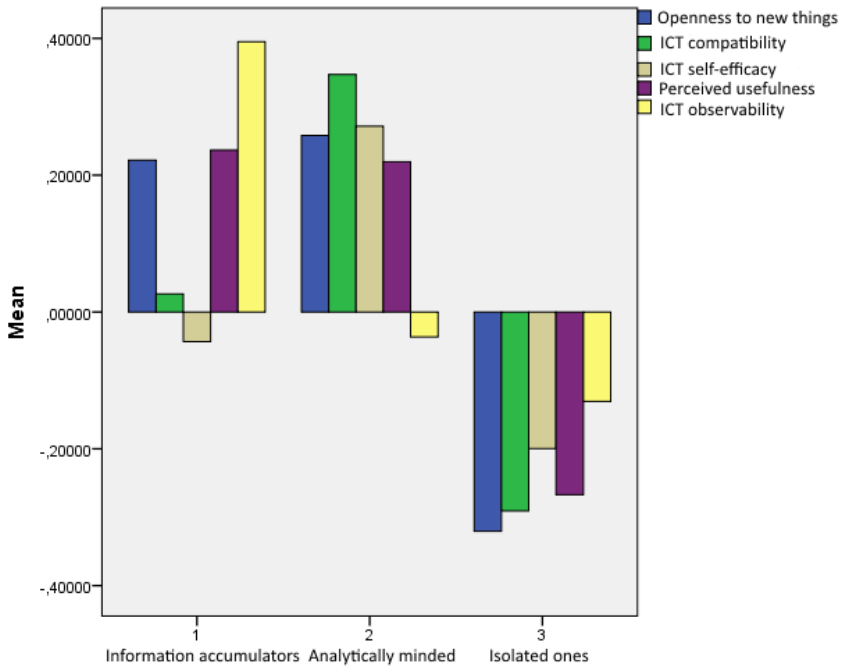

Figure 3. Latent variables within the clusters of farmers

The analytically-minded (group 2) are open to new things and agricultural ICT use is perfectly in line with their management style; moreover, they have good computer skills and are well aware of the benefits of ICT. Their agricultural qualifications help them to form an analytical way of thinking, thus significantly raising the use of agricultural-purpose software. The members of group 3 are typically closed to innovations, have little knowledge of ICT, nor do they see its advantages; consequently, ICT does not match their management style. This rather large third group mainly bear the traits of those members of the late majority or laggards, which can be seen clearly in the low adoption rates of smartphones in this group in comparison with the other two groups. The members of group 1 represent a kind of transition between the other two as they are aware of the advantages of ICT and the internet, they regularly experience the benefits of these technologies in their surroundings but their ICT skills (and confidence) are low, which probably prevents them from the agricultural use of ICT. Beyond the primary digital divide, a 'secondary agricultural digital divide' can be seen among farmers in regard to the use of agricultural-purpose ICT innovations. Group 1 holds special interest its members are innovative and open people who understand the benefits of ICT and the internet, frequently experiencing the beneficial effect of these in their surroundings; at the same time, they have poor ICT skills and little self-confidence in regard to ICT, which are likely to stop them from using ICT in agricultural activities, even though they have the opportunity to do so (however, since the proportion of those with small farms is the highest in this group, an increase in self-confidence would probably not automatically result in a sudden rise of software use). Based on the data regarding the various farmer groups, the research main hypothesis is proven.

4. Conclusion

Several conclusions can be drawn from the findings of this research. They can be successfully applied in the following areas: communication strategies with farmers, reducing the information deficit in the agricultural sector, designing ICT applications for farmers. It seems unambiguous that a significant group, amounting to close to one third of farmers in the study, has no openness to ICT innovations, its members not at all adapting these technologies, or even if they do, they do not exploit the potential benefits inherent in them – e.g. the numerous group of farmers who only use their mobile phones for conversations. A significant proportion of these people are not likely to use the most basic, general technologies in the near future. The likelihood of today's farm management support software being used by this group of farmers is negligible, since the intensity of their general-purpose ICT use and the self-confidence this would be coupled with do not reach the level which would enable the integration of such ICT solutions into daily farming activities. At the same time, information reaches the members of this group mainly via the general mass media: those organising agricultural applications and ICT solutions as well as the leaders in the agricultural sector must be aware of the specific needs and ways of reaching the members of this group. In the case of this group (about one third of farmers) intermediaries and agriculture extension workers will continue to play a great role in providing the mandatory transaction services and personalised information.

At the moment about one third of farmers (the most innovative third) fully and strategically exploit the benefits of ICT, even if the success of the farming log book software is partly explained by its mandatory nature for some EU-subsidies. These farmers practically already base their farm management activities on ICT; they actively gather information, use online transaction services and are open to using agricultural software. They can be the direct target groups and first users of new applications launched in the area, and they can be best reached at agricultural product presentations, fairs and via the agricultural press.

One quarter of farmers practically use ICT to the same extent as the previously mentioned group, they are still lagging behind in regard to agricultural ICT use, mainly because of their deficiencies in ICT knowledge and self-confidence, as well as the lack of an analytical way of thinking, these factors enhancing one another and resulting in a kind of 'secondary agricultural digital divide'. It is expected that with a relatively small investment this quarter of farmers can be turned into more active users if they are given sufficient support and the opportunity to try and use newly developed applications coupled with continuously available practical assistance. These efforts can be consolidated by increasing the self-confidence of these farmers as well as by clearly and transparently communicating

to them the advantages inherent in ICT – this is made easier by the fact that these farmers can be reached by a larger number of information channels. For those developing services and applications must clearly see that these farmers can be at best reached by solutions whose model and even user interface are 'hidden' behind a simple communication method (e.g. SMS consultation) in which knowledge transfer does not require farmers to adopt practices significantly different from the ones they are using.

It is an important conclusion of the research that agricultural qualifications have a significant impact on the use of different types of management software, this being a result of analytical thinking. The agricultural educational programmes should emphasise the use of online agricultural sources of information, applications and software from the start, and their integration into daily management tasks; thus, those open to ICT to start with can be orientated towards solutions aimed at enhancing efficiency.

References

Alvarez, J & Nuthall, P 2006. 'Adoption of computer based information systems: The case of dairy farmers in Canterbury, NZ, and Florida, Uruguay' *Computers and Electronics in Agriculture*, 50, pp.48–60. doi: 10.1016/j.compag.2005.08.013

Aubert, B, Schroeder, A & Grimaudo, J 2012. 'IT as Enabler of Sustainable Farming: An Empirical Analysis of Farmers' Adoption Decision of Precision Agriculture Technology' *Decision Support Systems* 54 (1), 510–520. doi:10.1016/j.dss.2012.07.002

Berman, A 2006. 'ICT in the Dairy Farming System' in ICT in Agriculture: Perspectives of Technological Innovation, eds E Gelb & A Offer, Samuel Neaman Institute for National Policy Research European Federation for Information Technologies in Agriculture, Food and the Environment (EFITA). http://departments.agri.huji.ac.il/economics/gelb-farming-6.pdf

Csótó, M 2003. 'Különbségek és azok feltárásának módjai a gazdálkodók információfogyasztásában és IKT-eszközhasználatában' (Exploring differences in farmers' information consumption and ICT-usage) *AGRÁRTUDOMÁNYI KÖZLEMÉNYEK = ACTA AGRARIA DEBRECENIENSIS* 52., pp. 91-98.

Doye, DG, Jolly, RW, Hornbaker, R, Cross, T, King, RP, Lazarus, W, Yeboah, A & Rister, E 2000. 'Farm Information Systems: Their Development and use in Decision-making', Ames, Iowa State University.

Gengyina, N 2009. 'The Concept of a Person's Information Culture: View from Russia' *HAL archives*. http://archivesic.ccsd.cnrs.fr/file/index/docid/359475/filename/TexteGendinaColloqueErte2008.pdf

Harkin, M 2006. 'ICT Adoption as an Agricultural Information Dissemination tool – an historical perspective' in ICT in Agriculture: Perspectives of Technological Innovation, eds E Gelb & A Offer, Samuel Neaman Institute for National Policy Research European Federation for Information Technologies in Agriculture, Food and the Environment (EFITA). http://departments.agri.huji.ac.il/economics/gelb-harkin-3.pdf

Herdon M, Csótó M, 2009. 'The Role of Intermediaries in the Success of Electronic Claiming for Farm Subsidies in Hungary' in 7th World Congress on Computers in Agriculture and Natural Resources, eds. FS Zazueta, J Xin, American Society of Agricultural Engineers, Michigan, pp. 117-120.

Hill, M 2009. 'Using farmer's information seeking behaviour to inform the design of extension' *Extension Farming Systems Journal*, 5(2), p.121–126. http://www.apen.org.au/efs-vol-5-no-2

Kozári, J 1994. 'A magyar mezőgazdaság ismereti és információs rendszer értékelése, fejlesztésének lehetséges irányai' (The evaluation and development opportunities for the Hungarian agricultural knowledge and information system) in Egységes Információs Rendszer alapjai a mezőgazdaságban Tudományos Tanácskozás, Gödöllő.

LaRose, R, DeMaagd, K, Chew, H, Tsai, H, Steinfield, C, Wildman, S & Bauer, J 2012. 'Measuring Sustainable Broadband Adoption: An Innovative Approach to Understanding Broadband Adoption and Use. International Journal of Communication 25 (6), pp. 2576–2600. http://ijoc.org/index.php/ijoc/article/view/1776/811

Molnár, Gy 2008. 'Az IKT-val támogatott tanulási környezet követelményei és fejlesztésilehetőségei' (Requirements and opportunities for the ICT-suported learning environment) Szakképzési Szemle 24 (3), pp. 257-278.

Öhlmér, B 1991. 'On-farm computers for farm management in Sweden: potentials and problems' *Agricultural Economics*, 5(3), pp.279–286. doi: 10.1016/0169-5150(91)90049-q

Řezník, T, Lukas, V, Charvát, K, Charvát Jr., K, Horáková, Š & Kepka, M 2016. 'FOODIE data models for precision agriculture' in Proceedings of the 13th International Conference on Precision Agriculture July 31 – August 4, 2016 St. Louis, Missouri, USA.

Sasvári, P. 2008. 'Az információs és kommunikációs technológia fejlettségének empirikus vizsgálata' (The development of information and communication technology; an empirical study) PhD thesis, Miskolc.

Solano, C, León, H, Pérez, E & Herrero, M 2003. 'The role of personal information sources on the decision-making process of Costa Rican dairy farmers' *Agricultural Systems*, 76(1), pp. 3–18. doi: 10.1016/S0308-521X(02)00074-4.

Sörensen, CG, Fountas, S, Nash, E, Pesonen, L, Bochtis, D, Pedersen, SM, Basso, B & Blackmore, SB 2010. 'Conceptual model of a future farm management information system ' *Computers and Electronics in Agriculture*, 72, pp.37–47. doi: 10.1016/j.compag.2010.02.003

Szabó, GG 2002. 'A szövetkezeti vertikális integráció fejlődése az élelmiszer-gazdaságban' (The development of cooperatives' vertical integration in the food industry) *Közgazdasági Szemle* XLIX/3. pp. 235–250.

Szabó, IL, 2000. *Családi gazdaságok és szövetkezeteik információs problémái a rendszerváltás után*. (Information-related problems of family farms and cooperatives after the regime change) Veszprémi Egyetem, Georgikon Mezőgazdaságtudományi Kar.

Szakál, F 1993. 'A családi gazdaságok szerepe a mezőgazdaság szerkezetében' (The role of the family farm in the structure of agriculture) *Gazdálkodás* 37 (7), pp. 1-9.

Vergot III, P, Israel, G & Mayo DE 2005. 'Sources and channels of information used by beef cattle producers in 12 counties of the Northwest Florida extension district' Journal of Extension 43(2) https://www.joe.org/joe/2005april/rb6.php

Z. Karvalics, L 2012. 'Információs kultúra, információs műveltség - egy fogalomcsalád értelme, terjedelme, tipológiája és története' (Information culture, information erudition – meaning, extension, tipology and history of a family of concepts) *Információs Társadalom*, 12(1), pp.7–43. http://www.infonia.hu/digitalis_folyoirat/2012/2012_1/i_tarsadalom_2012_1_karvalics.pdf

Hungarian Association of Agricultural Informatics
European Federation for Information Technology in Agriculture, Food and the Environment

Journal of Agricultural Informatics. Vol. 8, No. 1
journal.magisz.org

Comparison of Chi-square based algorithms for discretization of continuous chicken egg quality traits

Zeynel Cebeci[1], Figen Yildiz[2]

I N F O
Received 13 Nov. 2016
Accepted 21 Feb. 2017
Available on-line 15 Mar. 2017
Responsible Editor: M. Herdon

Keywords:
Data pre-processing, Supervised discretization, ChiMerge, Chi2, Extended Chi2, Modified Chi2.

A B S T R A C T

Discretization is a data pre-processing task for transforming continuous variables into discrete ones. In this study, four Chi-square based supervised discretization algorithms (ChiMerge, Chi2, Extended Chi2 and Modified Chi2) were compared for discretization of the fourteen continuous variables in a chicken egg quality traits dataset. We found that all of the algorithms had similar performances in term of training model accuracies obtained with C5.0 classification tree algorithm whereas ChiMerge and Chi2 were better than the remaining algorithms in term of training error rates. The numbers of intervals obtained with Chi2 tended to be large while they were very small in Extended Chi2 and Modified Chi2. The numbers of intervals from ChiMerge increased as the significance level increases whereas they were the same at all the levels of significance for the remaining algorithms. Consequently, it was revealed that ChiMerge at the significance levels of 0.05 and 0.10 was more efficient than the others and could be a better choice in discretization of the egg quality traits.

1. Introduction

Data mining is the collection of numerous methods and techniques to reveal meaningful patterns, valid and useful information in massive volumes of data. In many data mining applications such as feature selection, classification and association rules extraction, the majority of the algorithms have primarily been developed to run on discrete or categorical variables. On the other hand, the data are generally continuous and/or mixed type in many fields of study. Therefore, a discretization process is needed to turn continuous variables into discrete ones by splitting their range of values into a finite number of subranges called intervals, buckets or bins. As example of a continuous variable, the air temperature (^{o}C) can be transformed into three intervals as: (1) *low ($\leq$15)*, (2) *medium (16-29)*, (3) *high ($\geq$30)*. As in the temperature variable example, continuous variables are divided into finite numbers of intervals that are treated as categories by a discretization algorithm.The number of intervals produced in a discretization process is equal to the number of cut-points plus one. The minimum number of intervals for a continuous variable is equal to 1 while the maximum number of intervals is equal to the number of instances in a dataset.

In broad sense, a typical discretization consists of two stages. The first stage is a four-step task comprising of: (1) sorting values of continuous variables, (2) evaluating a cut-point for splitting or merging adjacent intervals, (3) splitting or merging intervals according to some criterion, and (4) stopping at some point depending on a termination criterion (Dash *et al.* 2011; Hemada & Lakshmi 2013; Kotsiantis & Kanellopoulos 2006; Liu *et al.* 2002). The second stage of discretization includes re-encoding all the values in the intervals. In this stage, each interval is labelled with a discrete value, and then the continuous values within an interval are mapped to the discrete value of corresponding interval. For discretization of continuous variables many discretization methods (or simply discretizers) had been developed. Although they are usually classified as supervised and unsupervised, they can also be classified in many different axes such as: (a) static versus dynamic, (b) local versus global, (c) bottom-

[1] Zeynel Cebeci
Div. of Biometry & Genetics, Faculty of Agriculture, Çukurova University, 01330 Adana - Turkey
zcebeci@cu.edu.tr
[2] Figen Yildiz
Div. of Biometry & Genetics, Faculty of Agriculture, Çukurova University, 01330 Adana - Turkey
yildizf@cu.edu.tr

up versus top-down, and (d) direct versus incremental (Kotsiantis & Kanellopoulos 2006; Ramírez-Gallego *et al.* 2015).

The supervised algorithms use priori known class labels information while unsupervised methods do not use such kind of information. In the static discretization algorithms, number of intervals is determined for each variable independently. Contrarily, the dynamic algorithms determine a possible number of intervals for all variables simultaneously. Since the multivariate algorithms capture interdependencies in discretization an overall improvement is expected in quality of discretization (Tay & Shen 2002). On the other hand, the static algorithms work for one variable at a time and thus they are synonymously called as the univariate algorithms. The dynamic algorithms are multivariate algorithms because they process multiple variables simultaneously. The local discretization algorithms use the local parts of instances space (subsets of instances) while the global algorithms run for the whole instances space (Chmelewski & Grzrmala-Busse 1996). The bottom-up algorithms (merging algorithms) are initialized with a complete list of all values as cut-points, and merge intervals by selecting the best cut-points. The top-down algorithms (splitting algorithms) start an empty list of cut-points and one interval covering all the values of a variable, and then divide this wide interval into smaller intervals with the best cut-points until a determined stopping criterion is reached. The direct algorithms for discretization need a user-defined number of intervals (*k* parameter) in discretization of continuous variables. In contrast to this disadvantage of the direct algorithms, the incremental algorithms do not require users to enter *k*. They start with a frontier discretization step and then search the best intervals in recursive improvements until a stopping criterion is satisfied.

Beyond it is necessarily needed by several data mining algorithms; discretization may also reduce the system memory requirement and shorten the execution time of the algorithms. Additionally, the information explored from discretized variables may be more compact and easily interpretable (Dash *et al.*, 2011; García *et al.* 2013; Gupta *et al*. 2010; Sang *et al*. 2013). In spite of its above mentioned advantages, discretization generally leads to certain level of information loss. Therefore, minimizing such loss is one of the main goals in developing discretization algorithms (García *et al*. 2013).

According to the surveys by Dougherty *et al*. (1995), Liu *et al*. (2002), Kotsiantis & Kanellopoulos (2006), García *et al.* (2013), and finally the advanced review by Ramírez-Gallego *et al*. (2015), many different discretization algorithms have been proposed in the last two decades. García *et al*. (2013) concluded that the most common techniques had been Equal-width Discretization (EWD) and Equal-frequency Discretization (EFD), MDLP, ID3, ChiMerge, 1R, D2, and Chi2. Among these, EWD and EFD are common unsupervised discretization methods due to their simplicity and availability in many data mining applications. However, they are direct algorithms that need an optimal *k* parameter (the number of intervals) for each variable before going to discretization process. Additionally, they have some other disadvantages such as having same values in different intervals and sensitivity to outliers. As an unsupervised alternative, although the K-means clustering algorithm overcomes the same value problem it is still sensitive to outliers. There is no superior algorithm for all of the data types yet the use of supervised algorithms may provide some advantages over unsupervised discretization, for istance they do not require user-defined parameters. On the other hand, the supervised algorithms have some disadvantages such as increase in time complexity which is a most common metric for measuring cost of an algorithm. For example, the time complexity is *O(n)* for EWD whereas it is *O(n log(n))* for ChiM based algorithms (Dash *et al*. 2011).

In the Chi-square based discretization algorithms, χ^2 statistic in Equation 1 is used to test the null hypothesis that two adjacent intervals are similar at a given significance level (α). When the adjacent intervals are independent they are merged, otherwise they are left separate.

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}} \qquad (1)$$

Where:

m : Number of intervals to be compared (usually *m=2*)

k : Number of classes

O_{ij}: Observed number of instances in i^{th} interval and j^{th} class

E_{ij}: Expected frequency of instances in i^{th} interval and j^{th} class $(= (r_i * c_j)\ /n)$

r_i : Number of instances in i^{th} interval ($= \sum_{j=1}^{k} O_{ij}$)
c_j : Number of instances in j^{th} class ($= \sum_{i=1}^{m} O_{ij}$)
n : Total number of instances ($= \sum_{j=1}^{k} c_j$)

In general, the Chi-square based algorithms are modifications or improved versions of ChiMerge (ChiM) algorithm. ChiM is a supervised and local merging algorithm applying the χ^2 testing in order to discretize continuous variables (Kerber, 1992). The algorithm performs discretization in two steps that are called the initialization and the bottom-up merging. In the initialization step, each distinct value of a continuous variable is assumed as an independent interval, and then χ^2 statistic is tested for whether the adjacent intervals to be merged or not. When the χ^2 statistic for adjacent intervals is greater than the predefined threshold level (χ^2_α), (if $\chi^2 > \chi^2_\alpha$), adjacent intervals are merged because they are assumed statistically similar. The χ^2 testing can be performed for adjacent interval pairs until a stopping criterion is satisfied.

Chi2 algorithm developed by Liu & Setiono (1995) is an extension of ChiM. This algorithm automates discretization by defining an inconsistency rate as stopping criterion and selects the statistical significance level automatically. It merges more adjacent intervals until the inconsistency criterion is satisfied. Modified Chi2 (mChi2) proposed by Tay & Shen (2002) is an improved modification of ChiM and Chi2. The mChi2 is a completely automatic algorithm fixing the over-merging problem because of the user-defined inconsistency rate which is leading to inaccuracy in Chi2. This algorithm uses a consistency rate from rough set theory to control inconsistency. Extended Chi2 (eChi2) is an extension of Chi2 in which inconsistency checking in Chi2 is replaced with a lowest upper bound (Su & Hsu 2005). In eChi2, two adjacent intervals is merged without considering difference with respect to degree of freedom. For this reason, eChi2 can handle uncertainty data, and obtained discretized data may result in better predictive accuracies when compared to those obtained from Chi2 and mChi2.

The StatDisc by Richeldi & Rossotto (1995) is an improvement of ChiM generating a discretization interval hierarchy by using the measurement as the interval merging criterion. Risvik (1997) proposed the Interval Merger technique which is a generalization of ChiM algorithm that decreases number of cut-points by removing each cut-point, and merging intervals until an inequivalent threshold value is achieved. The Concurrent Merger technique also uses the χ^2 statistic and inequivalence measures (Wang & Liu 1998). Khiops algorithm proposed by Boulle (2004) includes two steps which are the initialization step and the discretization optimization step. It differs from the previous Chi-square based algorithms with its stopping criterion rule and the use of global domain. Khiops does not require any pre-determined stopping criterion since it optimizes the χ^2 criterion in a global manner on entire instances space. Qu *et al*. (2008) proposed Rectified Chi2 algorithm aiming to fix the issues with mChi2 and eChi2 algorithms. Recently, Bettinger (2011) developed ChiD algorithm based on ChiM and Chi2.

In the literature, the researchers compared and proposed some discretization methods but they mostly worked with the classical benchmark datasets from the UCI Machine Learning Data Repository. So working with the real agricultural datasets is important in order to propose an appropriate method suitable for a specific domain in practice. For this reason, a comparative analysis has been given for discretization of 14 continuous variables in a chicken egg quality traits dataset in this study. Our aim was to generate a discretized dataset in order to use in a further study mining the association rules between these traits. Although there are many discretization algorithms, none of them are optimal for every situation. Based on the comparison of accuracy of PGN-classifier trained with different discretization methods on 8 datasets, Mitov *et al*. (2009) concluded that ChiM discretization method was more efficient for PGN-classifier than other methods. At the same time, as the examples of dynamic algorithms Chi-square based algorithms detect interdependencies between variables and discretize all variables concurrently. They are also known as the non-parametric algorithms which do not require any predefined parameter. On the basis of these advantages, we decided to compare some well-known Chi-square (χ^2) based algorithms for discretization of our dataset. In order to find a good algorithm for our purpose, we compared not only ChiM but also Chi2, eChi2 and mChi2 at the different levels of significance because of their availability in computing environments. In recent years, there are a few number of researches dealing with temporal data discretization for transforming the time series

into timely intervals (Azulay *et al*, 2007; Bakar *et al*. 2010; Acosta-Mesa *et al*. 2014; Chaudhari *et al*. 2014). In this study, we did not consider the temporal order of variables even they were measured weekly since ANOVA analyses showed that the majority of response variables did not differ by the time points of measurement (weeks).

2. Materials and Methods

In this study, we used an egg quality traits dataset containing various quantitative and qualitative variables recorded for totally 4320 eggs from the 3 commercial laying chicken lines (Lines A, D and N). The data was collected from a complete randomized plot design experiment conducted at the Experimental Farm of Faculty of Agriculture in Adana, Turkey. In the experiment, from each line 10 randomly sampled chickens were allocated to totally 18 cages in a three tiered (bottom, middle and top) and two sided (aisle and window) cages system, and raised for 24 weeks in a climate controlled poultry house. At the end of each week the eggs from each cage were collected and labelled for measuring the quality traits listed in Table 1.

In the analyzed dataset, there were 14 continuous variables and 1 class variable (genotype / line) as listed in Table 1. The dataset was checked and cleaned for the missing values and outliers before discretization. Firstly, all the data rows contain the missing values for at least 50% of variables were completely discarded from the dataset. The data size was reduced from 4320 to 4272 after this deletion. PMM (Predictive Mean Matching) imputation method was used in order to impute the remaining missing values in the analyzed dataset. Vink *et al*. (2014) stated that "PMM is very flexible as a method, because of its hot-deck characteristics, and is free of distributional assumptions. Moreover, PMM tends to preserve the distributions in the data, so the imputations remain close to the data". With respect to its above mentioned advantages, we applied PMM to our dataset by using the related functions of the `mice` package (van Buuren & Groothuis-Oudshoorn 2011) in R environment. Following the imputation of missing values, the records having the outliers below `Q1-1.5IQR` and above `Q3-1.5IQR` were successively discarded for each variable. The number of records was totally 3493 (Line A: 1146, Line D: 1187, Line N: 1146) after deletion of the outliers.

Table 1. Descriptive statistics for the continuous variables in the egg quality traits dataset

Vars	Description	Mean	SD	Min	Max	CV (%)	ADT (p)	#Outliers	#Intervals
ewg	Egg weight (g)	66.16	4.91	47.68	74.72	8.19	4.28e-10***	57	30
ewd	Egg width (mm)	43.19	1.22	42.38	46.67	2.83	2.00e-02*	62	33
eln	Egg length (mm)	56.93	2.23	50.43	63.52	3.92	1.04e-14***	44	32
eph	Egg pH	8.46	0.20	7.89	9.04	2.38	8.28e-06***	31	31
sbs	Shell breaking strength	4.68	1.05	1.70	7.65	22.44	1.20e-11***	119	31
sht	Shell thickness (µm)	366.40	22.64	303.33	429.43	6.18	1.40e-03**	42	31
shw	Shell weight (g)	6.80	0.64	4.99	8.66	9.48	1.73e-01^{ns}	45	31
ywg	Yolk weight (g)	16.08	1.90	11.03	21.23	11.80	2.80e-06***	42	31
yht	Yolk height (mm)	18.36	1.07	15.43	21.26	5.84	5.60e-01^{ns}	37	31
ywd	Yolk width (mm)	39.92	2.60	32.55	47.41	6.53	1.04e-06***	62	31
yce	Yolk color index (E)	81.77	5.36	66.29	97.42	6.55	3.00e-03***	74	31
wht	White height (mm)	8.64	1.15	5.32	11.78	13.32	1.10e-03**	39	31
wwd	White width (mm)	64.85	5.53	50.04	80.18	8.53	3.70e-24***	94	31
wln	White length (mm)	85.42	7.03	66.77	104.61	8.23	2.07e-10***	31	29
gen	Genotype of chicken	Class variable has three levels: A, D, N						-	-

* p<0.05, **p<0.01, ***p<0.001, ns: not significant

According to Kaoungku *et al*. (2015) non-normal distribution may have strong effect to discretization although it needs further theoretical and experimental proofs. In order to evaluate the probable effect of the distribution types on discretization performance we tested the normality of variables. Anderson-Darling Test (Anderson & Darling, 1952) is one of the preferred normality tests since it is more sensitive to deviations in the tails of the distribution and applicable for any type of data distribution. We used the `nortest` library (Gross & Ligges 2015) in R environment (R Core Team, 2015) for testing the normality of variables in our dataset. As seen from the ADT (p) column in Table 1, the variables shell weight (*shw*) and yolk height (*yht*) were normal and the remaining variables were non-normal ($p<0.05$). The coefficients of variation (CV% in Table 1) were 22.44%, 13.32% and 11.80% for the variables shell breaking strength (*sbs*), white height (*wht*) and yolk weight (*ywg*) respectively, and were under 10% and close to each other for the remaining variables.

In regard of a rough evaluation of the success of Chi-square based algorithms we aimed to compare the number of intervals generated by the studied algorithms to those of an unsupervised discretization method. For this aim we chose the Equal Width Discretization (EWD) as the representative of the unsupervised methods because of its simplicity and popularity in many studies. For EWD, however several rules such as Sturges', Scott's and Freedman-Diaconis do exist to obtain the interval width of variables ($h(x_i)$) Freedman-Diaconis rule (Freedman & Diaconis 1981) is one of the more informative rules due to inclusion of interquartile range statistic in calculation of interval widths as shown in Equation 2. Thus we generated the reference number of intervals by using EWD with the Freedman-Diaconis rule (EWI-FDR) in our study.

$$h(x_i) = 2\frac{IQR(x_i)}{\sqrt[3]{n}} \; ; k(x_i) = \frac{\max(x_i) - \min(x_i)}{h(x_i)} \qquad (2)$$

Where:
k: Number of intervals
h: Interval width
IQR: Interquartile range (*Q3-Q1*)
n: Number of instances

In this study, the original dataset was discretized by using the functions `chiM`, `chi2`, `extendChi2` and `modChi2` of the `discretization` library (Kim 2012) in the R statistical computing environment. Since the χ^2 test determines similarity of adjacent intervals based on the value of a statistical significance level (α), the levels of this parameter will affect number of intervals calculated. The χ^2 test is more conservative at the smaller significance levels, and thus less number of intervals is generated with the smaller significance levels. In general, researchers have used either the significance level of 0.01 or 0.05. For Chi2 algorithm, Kerber (1992) originally proposed to choose any of the significance levels of 0.01, 0.05 or 0.10. Yang *et al* (2011) used the significance levels of 0.01 and 0.05 for different types of datasets in their experiments. Although there is no a definite rule to choose an appropriate significance level (Liu & Setiono 1995), the use of smaller significance levels can be preferred in order to avoid excessive number of intervals. Hence, in addition to the commonly used significance levels we also included the significance level of 0.001 in order to see how this conservative level will affect the obtained number of intervals in our experiment.

In order to evaluate the discretization performances of the algorithms, we compared the number of intervals and the execution time required to discrete all of the variables in the analyzed dataset. In addition to these, the classification training error rates and test accuracies calculated with the C5.0 Decision Tree Algorithm were also used for comparing the performances of the algorithms. For calculating these values in the R environment we ran the `C5.0` function of `C50` library (Kuhn *et al.* 2015) on each discretized dataset. We defined the classification tree model by using all of the variables in Table 1 as the predictors (`X`) and the genotype of chickens (*gen*) as the class variable (`Y`), and ran the model for 10 iterations with boosting option. We randomly sampled 80% of the data points (*n*=2750) as the training dataset (`trainY` and `trainX`) and the remaining 20% (*n*=743) as the test dataset (`testY` and `testX`). The applied model was `C5.0` (`trainY ~., data = trainX, trials = 10`). All analysis were done on a PC with i7 processor, 16GB RAM and 1 TB HDD running under an x64 operating system.

3. Results and Discussion

As seen in Table 2, the numbers of intervals varied between 2-254 for ChiM, 17-126 for Chi2, 2-5 for eChi2 and 1-4 for mChi2. The numbers of intervals obtained from Chi2 were relatively larger while they were very small in eChi2 and mChi2. The numbers of intervals obtained with eChi2 and mChi2 were close to each other and smaller than those of EWI-FDR. Although the numbers of intervals from ChiM increased as the significance level increases, they remained the same at all levels of significance for the algorithms Chi2, eChi2 and mChi2. According to these findings, e-Chi2 and mChi2 could be considered not good because they produced few number of intervals which may not be sufficient to keep the information in continuous values of the variables. Moreover, mChi2 produced only 1 interval for the variables *yce* and *wwd*, and it can be considered as inefficient discretization because of production of the low number of intervals for almost all of the variables in the analyzed dataset.

For the normal distributed variables in the analyzed dataset such as *shw* and *yht*, ChiM at the 0.01 and 0.05 significance levels and Chi2 at all significance levels generated closer results to those of EWI-FDR. But similar trends were also observed for some of the non-normal variables such as *shw* and *sbs*. This comparison showed that the algorithms remained insensitive to distribution types for these variables. Similar evaluation was also valid for the variables with higher variation such as *sbs* and *wht* versus the variables with lower variation in the analyzed dataset. In this regard, we need the forthcoming studies to discover the effects of different distribution types and variability levels on discretization since variations of the variables in the analyzed dataset may be not enough for revealing probable effects of different levels of the variations.

Table 2. Number of the intervals by the studied algorithms at different levels of significance

		Significance levels (α)						Significance levels (α)			
Vars	Algorithms	0.001	0.01	0.05	0.10	Vars	Algorithms	0.001	0.01	0.05	0.10
ewg	ChiM	4	17	96	235	**ywg**	ChiM	5	22	64	126
	Chi2	96	96	96	96		Chi2	64	64	64	64
	eChi2	4	4	4	4		eChi2	5	5	4	5
	mChi2	4	4	4	4		mChi2	3	3	2	3
ewd	ChiM	3	6	42	92	**yht**	ChiM	2	9	44	80
	Chi2	42	42	42	42		Chi2	44	44	44	44
	eChi2	3	3	3	3		eChi2	2	2	2	2
	mChi2	3	3	3	3		mChi2	2	2	2	2
eln	ChiM	7	17	65	136	**ywd**	ChiM	4	19	61	138
	Chi2	65	65	65	65		Chi2	61	61	61	61
	eChi2	7	7	5	7		eChi2	4	4	2	4
	mChi2	3	3	3	3		mChi2	2	2	2	2
eph	ChiM	3	6	17	21	**yce**	ChiM	5	16	114	208
	Chi2	17	17	17	17		Chi2	114	114	114	114
	eChi2	3	3	3	3		eChi2	5	5	5	5
	mChi2	3	3	3	3		mChi2	1	1	1	1
sbs	ChiM	3	10	35	73	**wht**	ChiM	4	9	45	90
	Chi2	35	35	35	35		Chi2	45	45	45	45
	eChi2	3	3	3	3		eChi2	4	4	4	4
	mChi2	3	3	3	3		mChi2	3	3	3	3
sht	ChiM	5	11	30	54	**wwd**	ChiM	2	16	112	217
	Chi2	30	30	30	30		Chi2	112	112	112	112
	eChi2	5	5	4	5		eChi2	2	2	2	2
	mChi2	4	4	4	4		mChi2	1	1	1	1
shw	ChiM	5	11	23	48	**wln**	ChiM	2	16	126	254
	Chi2	23	23	23	23		Chi2	126	126	126	126
	eChi2	5	5	5	5		eChi2	2	2	2	2
	mChi2	4	4	4	4		mChi2	2	2	2	2

For each variable, the number of intervals from ChiM at the significance level of 0.05 was equal to those obtained from Chi2 at the significance level of 0.001 and the higher levels. This finding showed that ChiM at the significance level of 0.05 produced the same results with Chi2 at all significance levels. It was also interesting that, for all the variables, the numbers of intervals from ChiM at the significance level of 0.001 were equal to those obtained from eChi2 at all significance levels. Similarly the numbers of intervals from ChiM at the significance level of 0.05 were equal to those obtained from eChi2 at all significance levels. These findings showed that ChiM at significance levels of 0.001 and 0.05 produced the same results with eChi2 and Chi2 respectively. This is an important advantage in favor of ChiM when the cost of execution time is taken into account.

In discretized data, the intervals should keep the present information in the continuous values and not produce patterns so different from those in original dataset. Hence, the number of intervals from a discretization algorithm should not be too small or too large. As seen in Table 1, the number of intervals by the variables varied between 29 and 31 with EWD-FDR. Assuming these interval numbers are informative enough and regarding them as reference, ChiM at the significance levels of 0.01 and 0.05 produced the closest results to those from EWD-FDR for the majority of variables.

As seen in Table 3, the error rate of training model which was computed from the continuous values was 5.0%. When this error rate was used as the reference, the results showed that ChiM at the significance level of 0.001, eChi2 and mChi2 at all significance levels did not perform well enough because their training errors were 5-6 times bigger than those computed for continuous values. On the other hand, as seen in Figure 1, the training errors from ChiM at the significance levels of 0.05 and 0.10 were less than those computed for the continuous values in original dataset. Chi2 produced the similar results at all significance levels because it used same discretized datasets in all of them.

Table 3. The training error rates and the test accuracies of the training model by the algorithms

Dataset	Training Error (%)	Test Accuracy (%)
Original (Continuous)	5.0	53.2
ChiM-0.001	24.3	51.8
ChiM-0.01	5.6	49.2
ChiM-0.05	2.7	51.6
ChiM-0.10	2.4	55.0
Chi2-0.001	2.7	51.6
Chi2-0.01	2.7	51.6
Chi2-0.05	2.7	51.6
Chi2-0.10	2.7	51.6
eChi2-0.001	24.3	51.8
eChi2-0.01	24.3	51.8
eChi2-0.05	24.2	50.9
eChi2-0.10	24.3	51.8
mChi2-0.001	32.3	52.5
mChi2-0.01	32.3	52.5
mChi2-0.05	33.0	52.5
mChi2-0.10	32.3	52.5

Figure 1. The training error rates and the test accuracies of the training model by the algorithms

The test accuracy of the training model was 53.2% for the continuous values in original dataset. The training model resulted with a medium level of accuracy for all of the discretized datasets. As seen from Figure 1 and Table 3, the test accuracies computed on discretized datasets were nearly equal to each other and varied between 49.2% and 55.0%. The highest accuracy was obtained as 55.0% for ChiM at the significance level of 0.10. The smallest accuracy was 49.2% and again obtained from ChiM at the significance level of 0.01.

As seen from Table 4, eChi2 and mChi2 required more execution time because these algorithms are based on ChiM and Chi2. The longest execution time of 9.99 minutes was obtained for eChi2 at the significance level of 0.001. This algorithm also required longer execution time at the other significance levels. Excluding ChiM, the discretization time at the significance level of 0.001 was relatively longer when compared to the other levels of significance.

Table 4. Execution time (min) by the algorithms and the significance levels

	Significance levels (α)			
Algorithms	0.001	0.01	0.05	0.10
ChiM	8.56	8.80	8.52	8.44
Chi2	8.96	8.58	8.70	8.62
mChi2	9.34	9.02	9.32	9.33
eChi2	9.99	9.07	9.46	9.34

4. Conclusions

In comparison to the other algorithms, Chi2 generated larger numbers of intervals. Contrarily, eChi2 and mChi2 resulted with very small number of intervals. ChiM at the significance level of 0.01 produced more compatible results compared to those of EWD-FDR which was used as the reference unsupervised method.

Regarding the test accuracy of the training model there were no remarkable differences between the studied algorithms. On the other hand the training error rates were low for ChiM and Chi2 compared to those of eChi2 and mChi2. For the analyzed dataset, this result indicated that ChiM and Chi2 algorithms were better than eChi2 and mChi2. The number of intervals from ChiM at the significance level of 0.05 were equal to those obtained with Chi2 at the significance level of 0.001 for all of the variables. In addition to its acceptable performance in generation of the intervals, ChiM worked faster than Chi2, eChi2 and mChi2. As a consequence of these findings we recommend to work with ChiM at the significance levels of 0.05 or 0.10 for discretization of the chicken egg quality traits when the genotype is used as the class variable.

In this study, even though we compared four Chi-square based algorithms for nonparametric discretization of the continuous egg quality traits, the research still needs to consider with other families of the supervised discretization algorithms as well as the unsupervised methods. In future studies, for comparing the success of the algorithms we also plan to study on the other aspects of discretization such as to use the robustness or the accuracy criterion based on statistical tests.

Acknowledgement

We gratefully thank to Assoc. Prof. Dr. Mikail Baylan and his colleagues at the Çukurova University for their permission to use the dataset analyzed in this study.

References

Acosta-Mesa HG, Rechy-Ramírez F, Mezura-Montes E, Cruz-Ramírez N & Hernández JR (2014), 'Application of time series discretization using evolutionary programming for classification of precancerous cervical lesions', *J Biomedical Informatics*, vol. 49, p. 73-83. doi: 10.1016/j.jbi.2014.03.004

Anderson TW & Darling DA (1952), 'Asymptotic theory of certain 'goodness-of-fit' criteria based on stochastic processes', *Annals of Mathematical Statistics*, 23: 193–212. doi: 10.1214/aoms/1177729437

Azulay R, Moskovitch R, Stopel D, Verduijn M, de Jonge E & Shahar Y (2007), 'Temporal Discretization of medical time series - A comparative study', In *Working Notes of Intelligent Data Analysis in Biomedicine and Pharmacology*, July 8, 2007, p. 73-78.

Bakar AA, Ahmed AM & Hamdan AR (2010), 'Discretization of Time Series Dataset Using Relative Frequency and K-Nearest Neighbor Approach', in Advanced Data Mining and Applications, vol. 6440 of the series Lecture Notes in Computer Science, p. 193-201. doi: 10.1007/978-3-642-17316-5_18

Bettinger R (2011), 'ChiD, A χ^2-based discretization algorithm', *Proc. of Western Users of SAS Software*. San Francisco, California, US, October 12-14, 2011.

Boulle M (2004), 'Khiops: A statistical discretization method of continuous attributes', *Machine Learning*, vol. 55, no. 1, p. 53 – 69. doi: 10.1023/b:mach.0000019804.29836.05

Chaudhari P, Rana DP, Mehta RG, Mistry N & Raghuwanshi M (2014), 'Discretization of temporal data: A survey', *Int. J. of Computer Science and Information Security*, vol. 12, no. 2, p. 66-69.

Dash R, Paramguru RL & Dash R (2011), 'Comparative analysis of supervised and unsupervised discretization techniques', *Int. J. of Advances in Science and Technology*, vol. 2, no. 3, p. 29-37.

Dougherty J, Kohavi R & Sahami M (1995), 'Supervised and unsupervised discretization of continuous feature', In *Proc. of the 12th Int. Conf. on Machine Learning*, p. 194 – 202. doi: 10.1016/b978-1-55860-377-6.50032-3

Freedman D & Diaconis P (1981), 'On the histogram as a density estimator: L2 theory', *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 57, no. 4, p. 453 - 476.

García S, Luengo J, Sáez JA, López V & Herrera F (2013), 'Survey of discretization techniques, Taxonomy and empirical analysis in supervised learning', *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 4, p. 734 - 750. doi: https://doi.org/10.1109/tkde.2012.35

Gross J & Ligges U (2015), 'Nortest: Tests for normality', R package version 1.0-4, 2015. https://CRAN.R-project.org/package=nortest

Gupta A, Mehrotrab KG & Mohan C (2010), 'A clustering-based discretization for supervised learning', *Statistics and Probability Letters*, vol. 80, p. 816 – 824. doi: https://doi.org/10.1016/j.spl.2010.01.015

Hemada B & Lakshmi KSV (2013), 'A Study on discretization techniques', *Int. J. of Engineering Research and Technology*, vol. 2, no. 8, p. 1887-1892.

Kaoungku N, Thinsungnoen T, Durongdumrongchai P, Kerdprasop K & Kerdprasop N (2015), 'Discretization based on Chi2 algorithm and visualize technique for association rule mining', In *Proc. of the 3rd Int. Conf. on Industrial Application Engineering*, p. 254 - 260. doi: https://doi.org/10.12792/iciae2015.047

Kerber R (1992), 'ChiMerge: Discretization of numeric attribute', In *Proc. of the 10th National Conference on Artificial Intelligence*, p. 123 – 128.

Kim HJ (2012), 'Discretization: Data preprocessing, discretization for classification', R package version 1.0-1. (https://CRAN.R-project.org/package=discretization).

Kotsiantis S & Kanellopoulos D (2006), 'Discretization techniques: A recent survey', *GESTS Int. Transactions on Computer Science & Engineering*, vol. 32, no. 1, p. 47-58.

Kuhn M, Weston S, Coulter N & Clup M (2015), 'C50: C5.0 Decision Trees and Rule-Based Models', *R package version 0.1.0-24*. (C code for C5.0 by R. Quinlan License: GPL-3) (https://cran.r-project.org/web/packages/C50/).

Liu H, Hussain F, Tan CL & Dash M (2002), 'Discretization: An enabling technique', *Data Mining and Knowledge Discovery*, vol. 6, no. 4, p. 393 - 423.

Liu H & Setiono R (1995), 'Chi2: Feature selection and discretization of numeric attributes', *IEEE 24th Int. Conf. on Tools with Artificial Intelligence, IEEE Computer Society*, p. 388 – 388. doi: https://doi.org/10.1109/tai.1995.479783

Mitov I, Ivanova K, Markov K, Velychko V, Stanchev P & Vanhoof K (2009), "Comparison of discretization methods for preprocessing data for pyramidal growing network classification method". In *New Trends in Intelligent Technologies*, Int. Book Series Information Science & Computing - Book No: 142009, p. 31-39.

Qu W, Yan D, Sang Y, Liang H, Kitsuregawa M & Li K (2008), 'A novel Chi2 algorithm for discretization of continuous attributes', In *Proc. Progress in WWW Research and Development, 10th Asia-Pacific Web Conference*, China, April 26-28, 2009. p. 560 - 571. doi: https://doi.org/10.1007/978-3-540-78849-2_56

R Core Team (2015), 'R: A language and environment for statistical computing', R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ramírez-Gallego S, García S, Mouriño-Talín H, Martínez-Rego D, Bolón-Canedo V, Alonso-Betanzos A, Benítez JM & Herrera F (2015), 'Data discretization: taxonomy and big data challenge', *WIREs Data Mining Knowledge Discovery*, doi: 10.1002/widm.1173.

Richeldi M & Rossotto M (1995), 'Class-driven statistical discretization of continuous attributes' in *European Conference on Machine Learning*, p. 335 - 338. doi: https://doi.org/10.1007/3-540-59286-5_81

Risvik KM (1997), 'Discretization of numerical attributes: Preprocessing for machine learning', Computer Science Projects #45073, *Knowledge Sys. Grp., Dept. of Comp. and Inf. Sci. at Norwegian Univ. of Sci. and Tech. Trondheim, Norway*.

Sang Y, Zhu P, Li K, Qi H & Zhu Y (2013), 'A local and global discretization method', *Int. J. of Information Engineering*, vol. 3, no. 1, p. 6 – 17.

Su CT & Hsu JH (2005), 'An extended Chi2 algorithm for discretization of real value attributes', *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 3, p. 437 – 441. doi: https://doi.org/10.1109/icmlc.2012.6359019

Tay FEH & Shen L (2002), 'A modified Chi2 algorithm for discretization', *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 3, p. 666 – 670. doi: https://doi.org/10.1109/tkde.2002.1000349

Van Buuren S & Groothuis-Oudshoorn K (2011), 'mice: Multivariate imputation by chained equations in R', *Journal of Statistical Software*, vol. 45, no. 3, p. 1 - 67. http://www.jstatsoft.org/v45/i03/.

Vink G, Frank LE, Pannekoek J & van Buuren S (2014), 'Predictive mean matching imputation of semicontinuous variables', *Statistica Neerlandica*, vol. 68, no. 1, p. 61–90. doi: 10.1111/stan.12023

Wang K & Liu B (1998), 'Concurrent discretization of multiple attributes', *In the Pacific RIM Int. Conf. on Artificial Intelligence*, p. 250 – 259. doi: https://doi.org/10.1007/bfb0095274

Yang P, Li JS & Huang YX (2011), 'HDD: a hypercube division-based algorithm for discretisation', *Int. J. of Systems Science*, vol. 42, no. 4, p. 557–566. doi: https://doi.org/10.1080/00207720903572455

Hungarian Association of Agricultural Informatics
European Federation for Information Technology in Agriculture, Food and the Environment

Journal of Agricultural Informatics. Vol. 8, No. 1
journal.magisz.org

Special extension opportunities of CReMIT method in time series analysis and their application in forestry

Zoltán Pödör[1]

I N F O
Received 9 Jan. 2017
Accepted 6 Feb. 2017
Available on-line 15 Mar. 2017
Responsible Editor: M. Herdon

Keywords:
time series, CReMIT, climate, tree growth

A B S T R A C T

One of the problems frequently arising in connection with the study of time series is the thorough examination of relationships and correlations between the examined time series. In the case of time series with periodicity, study of the effects of periods with different time shifting, delayed, and varying length of windows can be also examined by using the early developed systematic CReMIT (Cyclic Reverse Moving Intervals Techniques) method. In this paper we present a special extension to this method that allows to alloy the CReMIT, the moving intervals and evolutionary techniques. It makes it possible to examine the temporal changes of the effects and relationships using evolutionary and moving interval techniques on special secondary time series derived by CReMIT. The function and applicability of the extended method are introduced on forestry, tree growth and meteorological data because of the weight of climate change. The climate change studies provide further motivation to examine the relationships between forestry parameters (tree growth) and environmental factors.

1. Introduction

Searching for relationships between time series is a major area of statistics and data mining. A huge number of techniques are available and among them correlation and regression analysis (Miles & Shevlin 2001; Myers 1990) are the most frequently used for defining connections between one or more independent and dependent variables. Beside the applied analysis methods, the completeness of the examinations can be significantly affected by the sphere of the involved dependent and independent variables. For example, the methods of cluster-analysis (Han & Kamber 2006), or PCA/ICA (Abdi & Williams 2010) can be adapted for the dimension reduction of datasets. In several steps of an analysis method, only a special portion of the full time series is used instead of the full range of the available time series. If we have a proper length time series, the temporal changes of the relationships can be examined by using the forward and backward evolution and moving interval techniques (Biondi & Waikul 2004). The essence of the moving interval technique is that the length of the examined interval is always fixed and the starting point is moved forward by one period in each iteration step. In the case of the evolutionary technique the starting point is fixed, and the interval length is increased by one period in each iteration step (Fritts 1970; Mudelsee 2010). These window-based techniques can be used not only during the breakdown of the whole examined data line into intervals, but also in a more specific manner in the case of periodic time series. Founded on the mentioned windows-based methods above, a special systematic window concept called CReMIT (Cyclic Reverse Moving Intervals Techniques) method combines the solution of moving intervals and evolutionary techniques, was created (Pödör, Edelényi & Jereb 2014a), This basic CReMIT makes it possible to systematically widen the sphere of the used dependent or independent variables with the derived transformed time series. It is important that the CReMIT method is independent of the applied analysis technique. However, the CReMIT uses the full, available time series to systematically generate the derived time series. The extension of the basic CReMIT method with the evolutionary and moving interval techniques insures the further extension of variables and the depth of the analysis. The extended CReMIT gives opportunities for the examination of the temporal changes of derived time series (based on the basic CReMIT) and relationships. The practical applicability of the extended

[1] Zoltán Pödör
University of Sopron
podor@inf.nyme.hu

CReMIT method is widely confirmed by the studies of correlations between yearly tree growth and monthly climate data. CReMIT was used on meteorological data, as independent variables to extend the analysis possibilities. The main aim of this paper is the description of the extended CReMIT method, but not the deduction of forestry results.

2. Short description of CReMIT

Let be ts a given time series ts and P be its natural period. The elements of this time series are stored in a vector. Let the first element of ts, ts_1 be the chronologically latest element, and natural numbers will be assigned to the data accordingly, the length of the vector is m.

$$(1)\ ts = \begin{pmatrix} ts_1 \\ ts_2 \\ \vdots \\ ts_m \end{pmatrix}$$

Let SP $(1 \le SP \le P)$ the SP^{th} element of the vector ts denotes the starting point of the currently applied investigation. Special windows are applied on the vector ts, the actual time shifting (i) and width (j) values of a window are defined on the basis of this index. The minimal value of time shifting can be 0 $(i = 0)$, and the window width can be 1 $(j = 0)$. Based on the period P of the basic time series, the above defined window will be periodically repeated with the maximum cycle number (MCN). As shown in Equation (2), the value of MCN depends on the defined parameters (SP, i, j):

$$(2)\ MCN_{ts} = \left[\frac{n - (SP + i + j)}{P} \right] + 1$$

, where $[\ ]$ is the entire function.

The starting and end point indexes of the windows created with the actual SP i and j values can be defined as $[SP + i + l * P; SP + i + j + l * P]$, where $0 \le l \le MCN - 1$. Using these parameters two temporal vectors are defined for the storage of the index values representing the limits of the windows as denoted in Equation (3) and Equation (4).

$$(3)\ index_{begin} = \begin{pmatrix} SP + i + 0 * P \\ SP + i + 1 * P \\ \vdots \\ SP + i + (MCN - 1) * P \end{pmatrix}$$

$$(4)\ index_{end} = \begin{pmatrix} SP + i + j + 0 * P \\ SP + i + j + 1 * P \\ \vdots \\ SP + i + j + (MCN - 1) * P \end{pmatrix}$$

By using the above defined index vectors a pre-defined transformation function TR (for example mean, maximum, minimum, sum) can be applied on the elements of the individual windows.

$$(5)\ tr\_ts_{SP,i,j} = \begin{pmatrix} TR(index_{begin}[1]; index_{end}[1]) \\ TR(index_{begin}[2]; index_{end}[2]) \\ \vdots \\ TR(index_{begin}[MCN]; index_{end}[MCN]) \end{pmatrix}$$

Based on the above described starting point SP $(1 \le SP \le P)$, the maximum time shifting value I $(0 \le i \le I)$ and the maximum window width J $(0 \le j \le J)$, pre-defined on the basis of the task all the potential $tr\_ts_{SP,i,j}$ transformed vectors can be generated on a systematic way. The values of I and J depend on the actual investigation and they are defined by the user.

2.1. Applicability of CReMIT

The CReMIT method was applied as part of an analytical process in order to ensure its application in practice. The process consists of three main parts: (a) data preparation, (b) execution of the CReMIT method, and (c) analytical modules. After data preparation is completed, the data enter the second transformation module that executes the CReMIT method. The third, the analytical module of the shell, receives the derived time series and executes the pre-defined analytical process. The derived time series generated by the CReMIT module can be used in any analytical process which ensures a high level of flexibility.

The practical applicability of the CReMIT method and the analytical process developed to expand the study of periodical time series are widely confirmed by the studies of correlations between tree growth, climate (Edelényi, Pödör, Jereb & Manninger 2011; Manninger, Edelényi, Pödör & Jereb 2011), and butterfly trapping data (Csóka, Pödör, Hirka, Führer & Szőcs 2012; Csóka, Pödör, Hirka, Führer, Móricz, Rasztovics & Szőcs 2013; Pödör, Csóka & Kiss 2013), the health of trees and climatic features (Pödör, Kolozs, Solti & Jereb 2014b).

The output size of the CReMIT method depends on the parametrization, and it is usually big. Therefore, the evaluation of the received results is really important for an expert from a special area of the studies.

3. The extension opportunities of CReMIT

The CReMIT method is independent of the applied analysis methods following the systematic extension of the basic time series, such as correlation and regression analysis (Myers 1990). However, the basic CReMIT method uses the full available, or the user selected ts time series to generate the $tr\_ts_{SP,i,j}$ transformed vectors. By using this approach, the relationships relative to the full ts can be examined. Nevertheless, sometimes we want to examine the temporal changes of the relationships, not only the static results for the full ts. It is especially important in the case of the environmental and forestry datasets - which defined the development of the basic CReMIT method – in view of the climate change.

3.1. The theoretical background

The extension of the basic CReMIT method with evolutionary (Figure 2) and moving interval (Figure 1) techniques makes it to possible to examine the temporal changes of the relationships based on ts. The evolution technique means, in fact, that the width of the applied window increases by one in each iteration without changing its starting point. In the case of moving intervals, the length of the examined interval is fixed in a suitable way and the starting point is moved forward by one cycle in every iteration step.

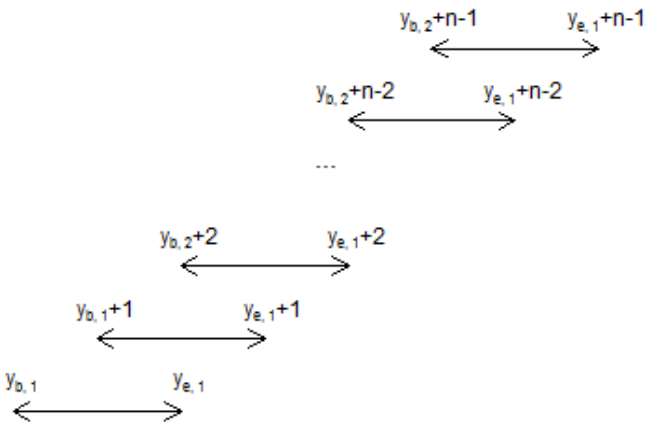

Figure 1. Moving interval technique for times series

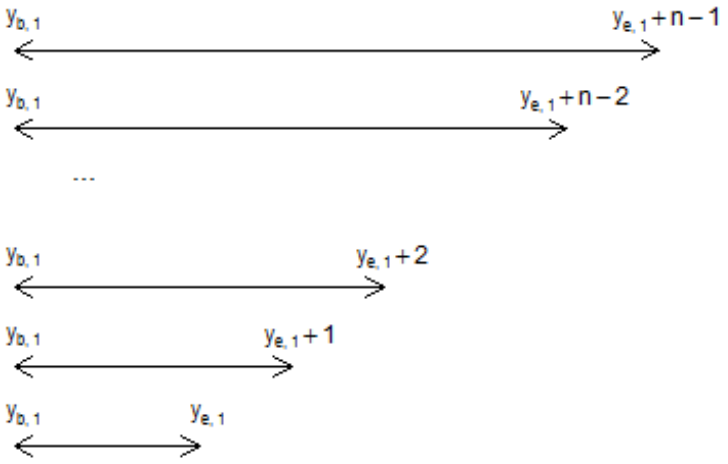

Figure 2. Evolutionary technique for time series

Let $tr\_ts_{SP,i,j}$, $(1 \le SP \le P)$, I $(0 \le i \le I)$, J $(0 \le j \le J)$, be the transformed vector corresponding to the basic CReMIT method with the parameters SP, i, j. The $tr\_ts_{SP,i,j}$ vectors are created on the whole length of the examined time series. Therefore, it is not suitable for the examination of the temporal changes of the relationships.

Let us suppose that the original ts time series contain m pieces of periods with length P (for example m is the number of years, and $P = 12$ by using monthly data in time series). The basic CReMIT method is complemented with the evolutionary and moving interval techniques. We are able to examine the temporal changes of relationships on given time shifting and window width. To do this we have to define a start and an end point. These points determine the initial interval of booth methods.

Let us denote the beginning period by BP, and the ending period by $EP < MCN$, where $EP - BP + 1$ is the initial interval length. The vector, representing the initial interval, comes into existence as shown in (6):

$$(6)\ tr\_ts_{SP,i,j}(1) = \begin{pmatrix} TR\left(index_{begin}[BP]; index_{end}[BP]\right) \\ TR\left(index_{begin}[BP+1]; index_{end}[BP+1]\right) \\ \vdots \\ TR\left(index_{begin}[EP]; index_{end}[EP]\right) \end{pmatrix}$$

Then, according to the selected procedure (moving interval, or evolutionary techniques), the additional intervals are generated. Let us denote by k the step kth of the given method. In the case of evolutionary technique, the length of the examined interval increases by one period step by step, without changing its starting point:

$$(7)\ tr\_ts_{SP,i,j}(k) = \begin{pmatrix} TR\left(index_{begin}[BP]; index_{end}[BP]\right) \\ TR\left(index_{begin}[BP+1]; index_{end}[BP+1]\right) \\ \vdots \\ TR\left(index_{begin}[EP+k-1]; index_{end}[EP+k-1]\right) \end{pmatrix}$$

By using the moving interval technique the length of the intervals are fixed, and the starting point is moved forward by one period step by step:

$$(8)\ tr\_ts_{SP,i,j}(k) = \begin{pmatrix} TR\left(index_{begin}[BP+k-1]; index_{end}[BP+k-1]\right) \\ TR\left(index_{begin}[BP+k]; index_{end}[BP+k]\right) \\ \vdots \\ TR\left(index_{begin}[EP+k-1]; index_{end}[EP+k-1]\right) \end{pmatrix}$$

The above described methods and the software frame were implemented in the open source R (R 2.15.2. version, R Development Core Team, 2008) environment and the mean, amount, minimum and maximum were the implemented elementary TR functions. The default analysing procedure was the linear correlation analysis with Student's t-test to evaluate the significance of the correlation coefficients.

3.2 Application of the extended method – an example

We represent the above described extension opportunities of the CReMIT method on a forestry dataset. Let the dependent parameter be the yearly tree growth data, and let the independent parameters be the monthly precipitation sum values. We would like to examine the relationships between the climatic factors (precipitation, temperature) and tree growth variables, but it is insufficient to compare only the simple monthly climatic and yearly tree growth time series.

The basic CReMIT method makes it possible to create and add several derived climatic variables - with different width and time shifting values for the examinations of the relationships. It has a really important part in the concern of the examined data sets; after all, the forestry parameter is affected by a period that is more than any given month. Moreover it is worthy to involve the climatic features of the past year in addition the actual year into the examinations.

At the same time, the basic CReMIT method uses the entire available time series in the process of the window creating. Therefore, it is unsuitable for the examination of the effect of a time interval – defined by a CReMIT created window – that is constant or that changes in time. It is an especially important task because of the effects of supposed climate change. Altogether it is uncertain that the same climatic variables of different time periods (for example: 1960-2010, 1960-1990, or 1990-2010) have the same effect on the same independent parameters. By using the extended CReMIT method we demonstrate the effect of changing climatic variables on the tree growth data.

In our example we examine yearly tree growth data (based on tree rings measurement) from 1962 to 2008, and monthly precipitation sum (from 1961 to 2008). The extended CReMIT method was used on the meteorological features (precipitation and temperature) as independent variables. The initialization values of the basic CReMIT method was defined by forestry professional:

$SP = 4$, September is the last examined month (in actual year),

$I = 12$, the maximal time shifting value,

$J = 5$, the maximal window width value.

In practice it means that all possible time windows were created from September of the actual year to April of the previous year with a maximum 6 month width for the monthly precipitation data. Altogether 93 different time periods (windows) were created. We used the undermentioned notations: *prec_x* shows that the examined parameter is the precipitation and *x* the number of month. The *a_*, and *p_* prefixes show that the data come from the previous, or from the actual year. For example *p_prec_10 - a_prec_3* means that the period from the previous year October to the actual year March; it is a window of 6 month width. Using the precipitation sum values of all created periods, it is possible to examine the effect of these periods on tree growth data.

But if we are curious about the temporal changes of the effects, then we should use the extended CReMIT method. Let the minimal interval length be 20 years, that is $BP = 1962$ and $EP = 1981$. By using this parametrization 28-28 intervals were generated by both the moving interval and the evolutionary techniques. Moving intervals: 1962-1981; 1963-1982; 1964-1983;…;1988-2007 and 1989-2008. Evolutionary technique intervals: 1962-1981; 1962-1982; 1962-1983;…; 1962-2007 and 1962-2008.

We analysed the relationships between the independent variables (93 time periods derived by extended CReMIT method, and 28-28 intervals) and the yearly tree growth data using simple linear correlation analysis. The significance of correlation coefficients were evaluated by Student's t-test at the level of significance α=0.05 with the number of degrees of freedom *n-2*. In Table 1 and Table 2 only the statistically significant correlation values were shown.

Table 1. The result of moving intervals and CReMIT method (details)

	BP	1962	1963	1964	…	1983	…	1988	1989
	EP	1981	1982	1983	…	2002	…	2007	2008
from	to								
p_prec_4	p_prec_4	0.53	0.56	0.68					
p_prec_4	p_prec_5		0.44	0.55					
…	…								
p_prec_10	p_prec_12	-0.39							
p_prec_10	a_prec_1	-0.44							
…	…								
a_prec_5	a_prec_9								
a_prec_6	a_prec_6								
a_prec_6	a_prec_7	0.4							
a_prec_6	a_prec_8					-0.38		-0.38	
a_prec_6	a_prec_9					-0.43			
a_prec_7	a_prec_7								
a_prec_7	a_prec_8					-0.38		-0.51	-0.44
a_prec_7	a_prec_9					-0.41			
a_prec_8	a_prec_8							-0.51	-0.55
a_prec_8	a_prec_9								
a_prec_9	a_prec_9								

Table 2. The result of evolutionary intervals and CReMIT method (details)

	BP	1962	1962	1962	…	1962	…	1962	1962
	EP	1981	1982	1983	…	2002	…	2007	2008
from	to								
p_prec_4	p_prec_4	0.53	0.51	0.5		0.25		0.22	0.23
p_prec_4	p_prec_5								
…	…								
p_prec_10	p_prec_12	-0.39	-0.39	-0.39		-0.25			
p_prec_10	a_prec_1	-0.44	-0.44	-0.44					
…	…								
a_prec_5	a_prec_9								
a_prec_6	a_prec_6					0.23		0.23	
a_prec_6	a_prec_7	0.4				0.24		0.26	
a_prec_6	a_prec_8								
a_prec_6	a_prec_9								
a_prec_7	a_prec_7								
a_prec_7	a_prec_8								
a_prec_7	a_prec_9								
a_prec_8	a_prec_8								
a_prec_8	a_prec_9					-0.21		-0.24	-0.24
a_prec_9	a_prec_9					-0.26			

The results in Table 1 and Table 2 showed that the relationships changed in time. There are time periods which have significant *r* values in the first 20 years periods while there are not in the middle, or the last periods, or vice versa. These result can be important in the course of the examination of the climate change effects.

Table 1 shows the results of 20 year long moving intervals (the starting points, BP are different) and Table 2 shows the results of different length time series (from 20 to 47 years), where the starting points (BP) are the same. The moving interval technique may represent the changing of the relationships better, because the starting points (BP) of the examined time intervals are moving in time. We can compare the relationships of the first 20 years (1962-1981) and the relationships of the last 20 years (1989-2008). It is important to detect the effect of climate change in nature.

In this chapter our aim was to introduce the applicability of the extended CReMIT method without declaring forestry results. Due to the size of the result tables, we only emphasize the parts which can show the change of relationships in time.

4. Conclusions

The study of time series and the analysis of their interrelations are highly important and are major areas of statistics and data mining and, therefore, the results presented herein may be related to these professional fields. The basic aims in analysing relationships are to define the strength of the relationships between variables and to define functions to describe these interrelations.

The efficiency of search for connections between time series can be affected by the applied analysis methods, the sphere of derived time series and the use of variables. A special time series transformation method based on CReMIT which can support the analysis of complex and deep relationships between the examined time series, is presented. The CReMIT method permits the systematic expansion of the parameters based on the applied window technique. It makes it possible to create different time periods beside the raw data (for example monthly data). These derived time series can be used in the relationships examinations.

The basic CReMIT method uses the full range of the examined time series, so it is not suited to the examination of the temporal changes of the relationships. A special expansion of CReMIT is given in this study which uses the evolutionary and the moving interval techniques to extend the sphere of

derived time series and the examination opportunities. This expansion gives us a real opportunity for the examination of the temporal changes of the relationships between the time series.

We illustrate the applicability and the function of this methodology through an example of forestry. The dependent variable is the yearly growth data of a tree from 1962 to 2008, and the basic independent variables are the monthly precipitation sums from 1961 to 2008. We can examine the effect of precipitation sums of different time periods from the previous year April to September of the actual year using the minimal interval length of 20 years. Due to the supposed climate change, it is an important area where we can efficiently use the presented methodology.

However, the applicability of the method is not limited to the forestry problems that had originally motivated its development the procedures developed can also be applied in any other field where the analysis of relationships between periodic time series is of fundamental importance.

References

Abdi, H & Williams, J L 2010 'Principal component analysis.' *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp 433–459.

Biondi, F & Waikul K 2004 'DENDROCLIM2002: A C++ program for statistical calibration of climate signals in tree-ring chronologies' *Comp Geosci* vol. 30, pp 303–311. doi: https://doi.org/10.1016/j.cageo.2003.11.004

Csóka, Gy, Pödör, Z, Hirka, A, Führer, E & Szőcs, L 2012 'Influence of weather conditions on population fluctuations of the oak processionary moth (*Thaumetopoea processionea* L.) in Hungary' *Joint IUFRO 7.03.10 – "Methodology of forest insect and disease survey" and IUFRO WP 7.03.06 – "Integrated management of forest defoliating insects" Working Party Meeting*, 10-14 Sep 2012, Palanga, pp 10.

Csóka, Gy, Pödör, Z, Hirka, A, Führer, E, Móricz, N, Rasztovics, E & Szőcs L 2013 'Időjárásfüggő fluktuáció a tölgy búcsújáró lepke nyugat-magyarországi populációinál' *Növényvédelmi Tudományos Napok* 19-20 Feb 2013, Budapest

Edelényi, M, Pödör, Z, Jereb, L & Manninger M 2011 'Developing of method for transforming time series data and demonstration on forestry data (in Hungarian)' Acta Agraria Kaposváriensis vol. 15, no. 3, pp 39-49.

Fritts, H C 1976 Tree rings and climate Academic Press, London, p 582.

Han, J & Kamber, M 2006 'Data Mining: Concepts and Techniques, 2nd ed.', Morgan Kaufmann Publishers, p 703.

Mudelsee M 2010 Climate Time Series Analysis – Classical Statistical and Bootstrap Methods, Springer, p 467.

Manninger, M, Edelényi, M, Pödör, Z & Jereb L 2011 'The effect of temperature and precipitation on growth of beech (*Fagus sylvatica* L.*)* in Mátra Mountains, Hungary' *Applied Forestry Research in the 21st Century conference*, 13-15 Sep 2011, Prága-Pruhonice pp 22-23.

Miles, J & Shevlin, M 2001 'Applying Regression and Correlation: A Guide for Students and Researchers' Sage publications Ltd: p 251.

Myers H R 1990 'Classical and modern regression with applications (second edition)' Virginia Polytechnic Institute and State University, Duxbury, Thomson Learning: p 488.

Pödör, Z, Csóka, Gy & Kiss B 2013 'Simple- and Multivariate data analysis of light trap catching data by a systematic window procedure' *Decision Support System Workshop and ForestDSS Community of Practice*, 4-6 Dec 2013, Lisbon

Pödör, Z, Edelényi, M & Jereb L 2014a 'Systematic Analysis of Time Series – CReMIT' *Infocommunication Journal*, vol. 6, no. 1, pp 16-22.

Pödör, Z, Kolozs, Solti Gy & Jereb L 2014b 'Investigation of Hungarian forest health condition with special respect to climate change' *Journal of advances in agriculture* vol. 3, no. 2, pp 164-176.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Hungarian Association of Agricultural Informatics
European Federation for Information Technology in Agriculture, Food and the Environment

Journal of Agricultural Informatics. Vol. 8, No. 1
journal.magisz.org

Assessment of the milking machine parameters using a computer driven test system

Roşca Radu[1], Ţenu Ioan[2], Cârlescu Petru[3]

I N F O
Received 22 Aug. 2016
Accepted 5 Dec. 2016
Available on-line 15 Mar. 2017
Responsible Editor: M. Herdon

Keywords:
mechanical milking, teatcup, pressure sensor, force sensor, liner.

A B S T R A C T

The purpose of the paper is the development and evaluation of a computer driven system for the assessment of the mechanical milking machines. The tests were performed with the WestfaliaSurge Classic Pro liner and the WestfaliaSurge Classic liner, at pulsation rates of 50, 55 and 60 cycles/min and pulsation ratios of 60/40 and 50/50, at a vacuum level of 40 kPa (61.3 kPa absolute pressure). The recorded data was used to evaluate the durations of the pulsation phases, the teat-liner contact pressure, the pressure difference at which the liners starts to close and the time the liner is open and respectively closed. The experimental results confirmed the functionality of the system; differences were recorded between the theoretical (set) pulsation rates and the real ones (which are lower); for most of the regimes (10 of the 12 tested), the relative differences did not exceed 5%. As far as the pulsation ratio was concerned, the relative differences between the values prescribed by the computer software and the recorded ones were lower than 3% for 10 of the 12 variants. The results concerning the other pulsation characteristics were in accordance with the results reported by other authors, with the lowest contact pressures being recorded for the ClassicPro liner.

1. Introduction

The principle of mechanic milking is based on the pressure difference between the udder and the vacuum applied to the teat. In order to limit the development of congestion and edema and provide relief to the teat from the milking vacuum, the pulsation principle is used (Mein, Williams, Thiel, 1987); the ISO 3918: 2007 standard defines pulsation as the cyclic opening and closing of the teatcup liner. Collapse of the teatcup liner beneath the teat is achieved when air at atmospheric pressure is admitted into the pulsation chamber of the teatcup; the liner opens, allowing the extraction of milk, when vacuum is applied to the pulsation chamber. Figure 1 presents the typical pulsation cycle and phases, defined according to the ISO 5707: 2007: a is the increasing vacuum phase, b is the maximum vacuum phase, c is the decreasing vacuum phase and d is the minimum vacuum phase. The pulsation ratio is defined as the ratio between the duration of the a + b phases and the duration of the entire cycle.

According to Bade, Reinemann, Zucali, Ruegg & Thompson (2009), the pulsation rate and ratio, the vacuum level and the compressive load applied to the teat when the liner collapses are the factors affecting the peak milk flow rate: the flow rate increases when the vacuum applied to the teat end and the duration of the b phase increases; in the meantime, the liner compression should increase in order

[1] Roşca Radu
University of Agricultural Sciences and Veterinary Medicine Iaşi
rrosca@uaiasi.ro
[2] Ţenu Ioan
University of Agricultural Sciences and Veterinary Medicine Iaşi
itenu@uaiasi.ro
[3] Cârlescu Petru
University of Agricultural Sciences and Veterinary Medicine Iaşi
pcarlescu@yahoo.com

to relieve tissue congestion due to the higher milking vacuum. Adley & Butler (1994) stated that inadequate liner collapse could lead to high infection levels. Mein & Reinemann (2009) also concluded that the liner compression should increase when the milking vacuum is increased, but also mentioned that an increased liner compression has a negative effect over the teat-end condition, leading to the development of teat-end hyperkeratosis; they also showed that an increased duration of the b phase led to a higher peak milk flow rate.

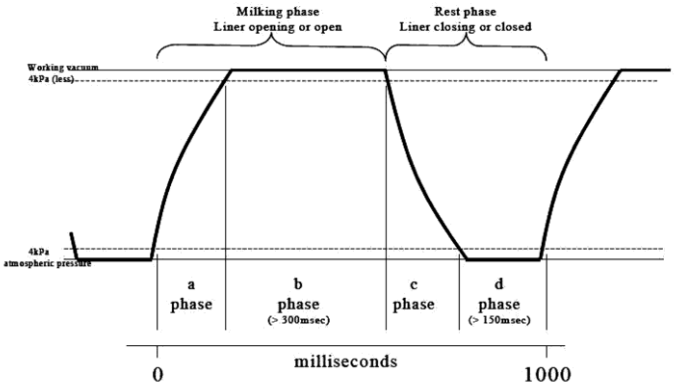

Figure 1. The pulsation cycle[4]

In a paper presented at the NMC 47th Annual Meeting by Kochman, Laney & Spencer the importance of the c phase was emphasized; the authors stated that, for a shorter c phase, the increased closing speed of the liner could cause physical discomfort to the cow, with negative results over the milking performance.

Billon & Gaudin (2001) showed that the milking time and milk flow rates were affected by the duration of the a and c phases; shorter phases led to longer milking times and lower flow rates.

All these facts emphasize the importance of developing measuring systems in order to evaluate the vacuum level, the phases of the pulsation cycle and the liner-teat contact pressure. The ISO 6690:2006 standard imposes a minimum sample rate of 100 Hz for the test of pulsators.

Reinemann, Rasmussen & Mein (2001) used Px139 low cost, amplified output, absolute pressure transducers in order to measure the vacuum levels in milking machines. These sensors used the bending membrane principle for measuring pressure; a thin conductive layer was applied over the membrane and the resistance of the conductive layer changed when the membrane was bended. The response time of the sensors was less than one ms. In order to evaluate the pulsation characteristics, the pressure sensor was connected to the short pulse tube. The vacuum level was measured at the teat end and claw. The authors concluded that, because the claw is easy to access, for routine field investigations the milking vacuum should be measured here; for measuring the vacuum in the short milk tube, the vacuum sensor should be located very near to the measuring point. Sensors based on the same principle (MICRO SWITCH 141PC15GL, 1 ms response time) were used by Spencer & Jones (2000) for measuring vacuum in the milking system.

The measurement of the liner-teat contact pressure rises at least two problems: what type of teat (artificial or live excised teat) and what type of pressure transducer should be used. Davis, Reinemann & Mein (2001) used an excised teat tip and a load cell in order to measure the compressive load over a sensor with different coverings; Reinemann, Mein & Muthukumarappan (1994) used an extruded clay ribbon, a flat tube and an excised teat and also an artificial teat equipped with a pressure sensor in order to measure the compressive load applied by the collapsed liner; Adley & Butler (1994) also developed and artificial teat, equipped with a load cell, loaded by a free piston and covered with a thin latex tube; van der Toll, Schrader & Aernouts (2010) showed that artificial teat was subjected to higher compression loads compared with an actual teat, because of the rigidity of the former. From this point of view, excised live teats are expected to provide more realistic results when compared to

[4] www.cowtime.com.au/technical/QuickNotes/Quick_Note_4_3.pdf

the artificial ones, because it is rather difficult to extrapolate the results from an artificial teat to a real one; on the other hand, even when an excised teat is used, due to the large variety of existing animals, it is also difficult to extrapolate the results obtained in this manner; moreover, in the every day life, the use of an excised teat can not be regarded as a common practice for field testing of the milking machines.

It is obvious that both these methods cannot be applied to live animals; in this case only the use of a specially prepared liner, containing a pressure sensor can be taken into account; Gates and Scott (1986) measured the compressive load with the help of a pressure transducer mounted inside the wall of the liner.

Referring to the pressure transducer, many researches consider that the maximum contact pressure is applied to the apex of the teat, but it is difficult to mount a pressure sensor in this area, so that the general practice is to place the sensor on the lateral surface of the teat. Mainly two types of devices were considered for measuring the liner-teat contact pressure: load cells and transducers based on a flexible pressure-sensitive layer. For the first type of sensor, the load cell is placed inside an artificial teat (covered with latex – Adley & Butler, 1994 - or even with an excised artificial teat – Davis, Reinemann & Mein, 2001) and the pressure from the liner-teat interface is applied to the load cell by the means of a circular piston (Adley & Butler, 1994) or of different sensor coverings (Davis, Reinemann & Mein, 2001). The friction forces between the piston and its bore affect the compressive load transmitted to the load cell; the relatively significant cost of the load cell is also a disadvantage. For the second type of sensors, while some authors (Reinemann, Mein & Muthukumarappan, 1994) concluded that the use of a flexible pressure-sensitive layer is not an accurate measuring method, others, like van der Toll, Schrader & Aernouts (2010), used it in order to measure the pressure at the teat-liner interface and concluded that the horizontal shear forces did not degrade the sensor's pressure readings. This method has some advantages: the sensor is easily applied on the surface of the artificial teat and does not disturb the pressure distribution because it is thin; the sensors are relatively cheap; the signal conditioning circuit is a very simple one.

Demba, Elsholz, Ammon & Rose-Meierhöfer (2016) used pressure-indicating films in order to measure the contact pressure distribution. Under pressure red patches appear on the film and the density of the color indicates the pressure level. The films were applied on both rigid and flexibile artificial teats. A dedicated software was then used in order to analyze the pressure distribution. The authors reported significantly higher contact pressures (70 to 640 kPa) than those found in other investigations.

It is obvious that some of the above-mentioned methods (live excised teat, liner with pressure transducer, pressure-indicating films) are difficult to apply outside an adequate research laboratory. Taking into account this matter, the general objective of the study is to present and evaluate a relatively cheap computer driven test system, which can be used outside a specialized test laboratory in order to measure the working parameters of the mechanical milking system. The system may also be used in field conditions in order to study the effect of the pulsation characteristics (rate and ratio) over the cows' health and milk yield.

2. DEVELOPMENT OF THE TEST SYSTEM

The developed system consists of two parts:

- a computer controlled impulses generator, completed with an electromagnetic pulsator, which can be used to test the teatcup liner in different working conditions (pulsation rates and ratios);

- a pressure recording system, used for monitoring the claw vacuum, the short pulse tube vacuum and the liner-teat contact pressure. An artificial teat, according to the specifications of the ISO 6690 standard, equipped with a force transducer, was used for recording the contact pressure between the collapsed teatcup liner and the artificial teat.

2.1. The computer controlled impulses generator

The computer controlled impulses generator is driving the electromagnetic pulsator by the means

of the computer software and electronic hardware; a slider valve type electromagnetic pulsator was used in this study.

The electronic hardware consists of a control board with 14 ports, produced by PC Control Ltd.[5], connected to the USB port of the computer; only the first two ports were used for the command of the electromagnetic pulsator.

The computer software is written in Visual Basic 6 and allows the adjustment of the pulsation rate and ratio; the flowchart of the program is presented in Figure 2. The program uses a dynamic link library (mb.dll), which encapsulates the functions needed for the communication with the control board across the USB interface.

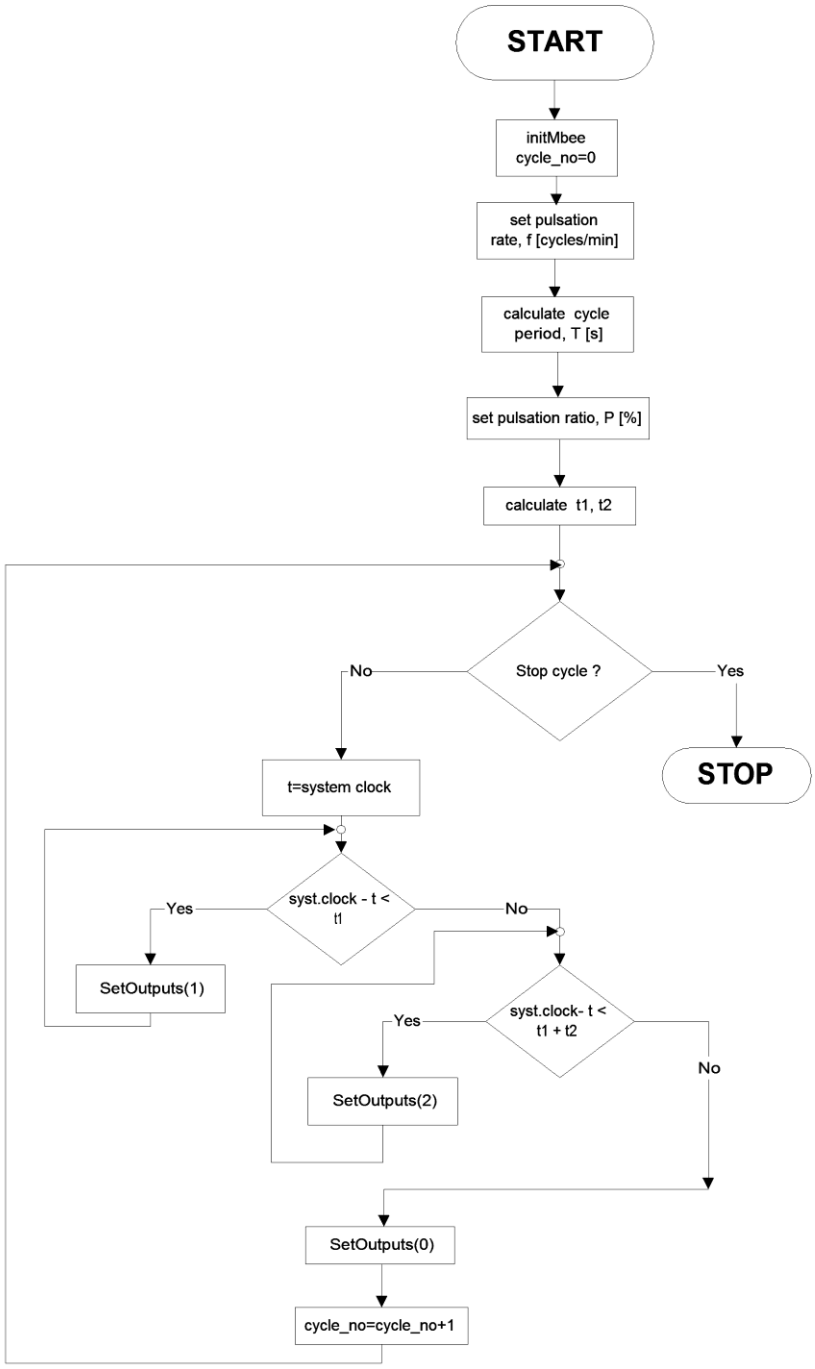

Figure 2. Flowchart of the computer program

[5] http://www.pc-control.co.uk/minibee_info.htm

As shown in Figure 2, the program uses the system clock in order to calculate the milk extraction time and the duration of the rest phase, starting from the set pulsation rate and ratio. The graphical user interface (GUI, Figure 3) of the program is used to set the cycle rate f [cycles/min] and the pulsation ratio, P [%]. The pulsation ratio may be adjusted between 10 and 90% and the pulsation rate may be adjusted in the range 10 – 120 cycles/min.

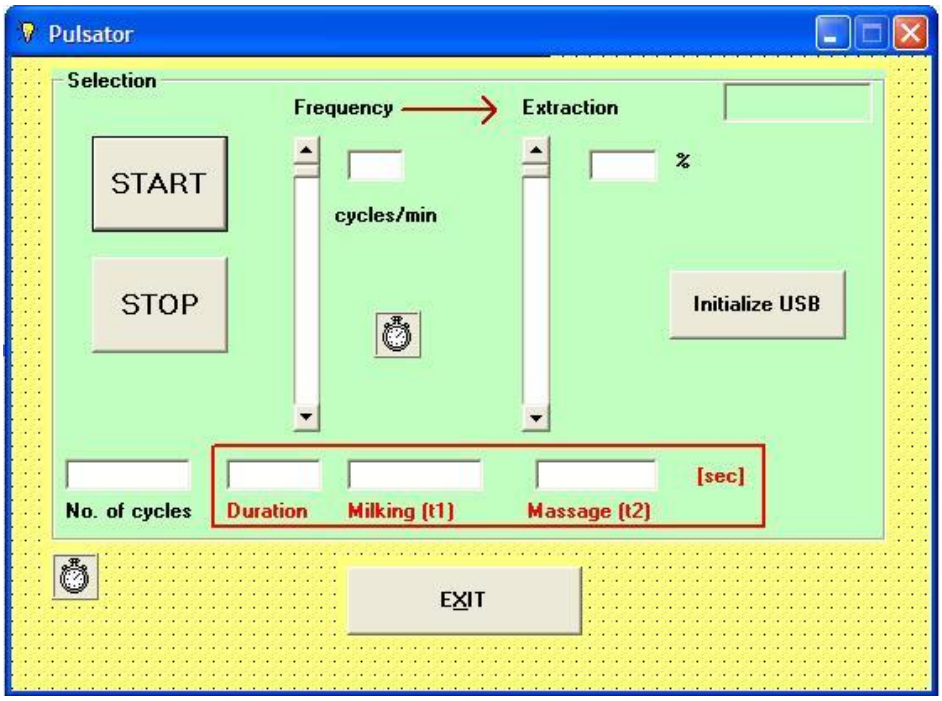

Figure 3. The GUI of the computer program

It should be mentioned that the pulsation ratio defined here is the ratio between the time while the coil of the pulsator is energized and the pulsation cycle duration, T; as it will be shown later, differences were recorded between this value (also called pulsator ratio) and the one defined according to the ISO 5707:2007 standard, which takes into account the intermittent vacuum signal recorded into the pulsation chamber or short pulse tube.

2.2. The pressure recording system

The computer controlled pressure recording system consists of:

- absolute pressure transducers type SPD015AAsil (SMARTEC[6]), with analogical output and the absolute pressure range between 15 and 102 kPa, which are used for sensing the claw vacuum inside and the intermittent vacuum into the short pulse tube of the teatcup. The sensors are based on the same principle as the ones used by Reinemann, Rasmussen & Mein (2001) and have the same response time – 1 ms.

- a data acquision board type USB6009 (National Instruments), with a sample rate of 48 ksamples/s and 4 differential analog input channels.

- an artificial teat, manufactured according to the specifications of the ISO 6690 standard (Figure 4), equipped with an A201 FlexiForce (Tekscan) type force transducer, with a diameter of the sensing area of 9.53 mm.

- a virtual instrument, designed with the LabView 7.1 software package, allowing both the visualization and the recording of the pressure signals.

Because force sensing is based on the modification of the electrical resistance of the transducer, a signal conditioning circuit was used in order to convert the variation of the electrical resistance into a

[6] http://www.smartec-sensors.com/assets/files/pdf/Datasheets_pressure_sensors/SPD015AAsilN.pdf

voltage signal. Bending of the force sensor over the round surface of the teat did not affect the readings (0 volts at output of the conditioning circuit when there was no force applied).

The block diagram of the virtual instrument created with LabView is presented in Figure 5; it uses a DAQmx virtual channel in order to read two physical channels (ai:0 and ai:1) of the USB 6009 device. The virtual channel measures voltages between zero and five volts; the differential terminal configuration (code 10106) is used for each physical channel. The sample clock allows the adjustment of the sampling rate, in samples per channel per second; during the tests, the sampling rate was set at 100 samples/s. A waveform chart was used to display the pressure signals on the computer screen; in the meantime, the signals were saved in a computer file.

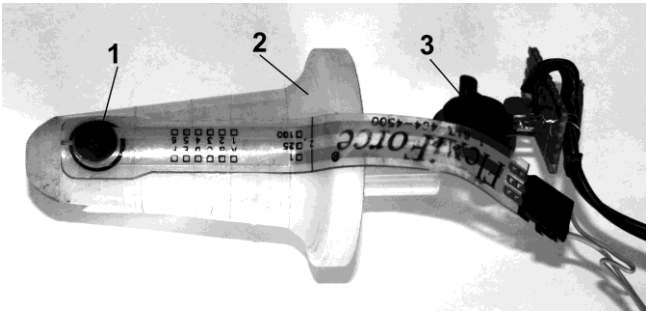

Figure 4. View of the artificial teat:
1-force sensor; 2-artificial teat; 3-absolute pressure sensor

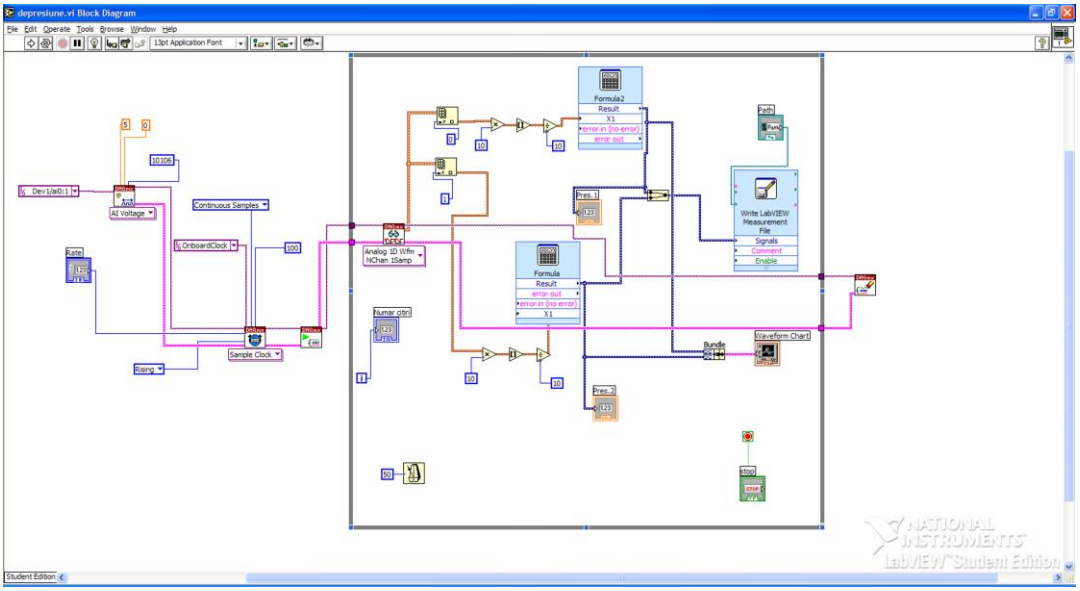

Figure 5. Block diagram of the virtual instrument

3. Assessment of the test system

In order to evaluate the system two types of teatcups were tested:

- WestfaliaSurge Classic Pro silicone liner, part. no 7029-2725-010, for cow teats with the diameter between 21 and 28 mm, a mouthpiece diameter of 23 mm and the wall thickness of 2.5 mm; when mounted into the shell, the teatcup liner is elongated by 6.8%.
- WestfaliaSurge Classic rubber liner, part no. 7021-2725-350, for cow teats with the diameter between 20 and 27 mm, a mouthpiece diameter of 23 mm and the wall thickness of 2 mm; when mounted into the shell, the teatcup liner is elongated by 5.4%.

A low milk line type milking machine was used, the vacuum level in the main airline was set at of 40 kPa (61.32 kPa absolute pressure); the permanent vacuum level measured under the teat (inside the liner) was 38 kPa (63.32 kPa absolute pressure).

The artificial teat was mounted into the teatcup taking into account the shape of the collapsed liner, so that maximum contact pressure was applied upon the area where the force sensor was placed.

Two pulsation ratios (60/40 and 50/50) and three pulsation rates (50, 55 and 60 cycles/min) were used during the tests.

As shown by van der Toll, Schrader & Aernouts (2010) and in order to allow the liners to slide more readily around the teat, a lubricant was sprayed onto the surface of the artificial teat.

In order to obtain a high degree of repeatability for the same liner, the artificial teat was left undisturbed between the experiments (Adley & Butler, 1994).

Before performing any measurement, the liners were pulsated for at least 50 minutes; there were no significant variations of the environmental temperature during the experiments.

For each test, the following parameters were evaluated:
- the pulsation ratio and rate;
- the duration of the cycle phases;
- the time the teatcup liner is completely open;
- the time the liner is closed;
- the maximum contact pressure between the liner and the artificial teat;
- the critical collapsing pressure difference.

The phases of the pulsation cycle were defined according to the requirements of the ISO 5707:2007 standard, using the pressure signal from the short pulse tube. The permanent vacuum was measured inside the liner, by the means of a channel drilled along the axis of the artificial teatcup

The pulsation cycle ratio and rate were calculated with the relations:

$$P = \frac{t_a + t_b}{T} \cdot 100 \quad [\%], \qquad (1)$$

$$f = \frac{60}{T} \quad [cycles/\,min], \qquad (2)$$

where $T = t_a + t_b + t_c + t_d$ is the cycle period [s], and t_a, t_b, t_c and t_d are the durations of the respective phases [s]. The relative differences between the set values of the pulsation rate and ratio and the measured ones were then calculated.

The force sensor mounted on the artificial teat was also used in order to evaluate the time the teatcup was completely open, τ_o (Figure 6), considering the time when no force was applied to the teat. The time the liner was closed was considered as the sum of the durations for the c and d phases.

The maximum contact pressure between the teatcup liner and the artificial teat was measured during the d phase of the pulsation cycle.

The critical collapsing pressure difference Δp was defined as the point at which the liner began to buckle because of the pressure difference across the walls, applying pressure over the teat (during the c phase of the pulsation cycle) and was evaluated with reference to the liner absolute pressure, as shown in Figure 6.

The critical collapsing pressure difference was measured for each type of liner and operating condition, after the liners were pulsated for about 50 minutes.

For each working condition (pulsation rate, pulsation ratio and type of teatcup) at least three tests were performed and then the average values were calculated.

For the same pulsation ratio an analysis of variance (ANOVA) was applied with respect to each of

the measured parameters in order to find out if the results were significantly different from a liner to another ($p<0.05$). The conclusion was that, except for the durations of the a and c phases, all the tested variants were significantly different.

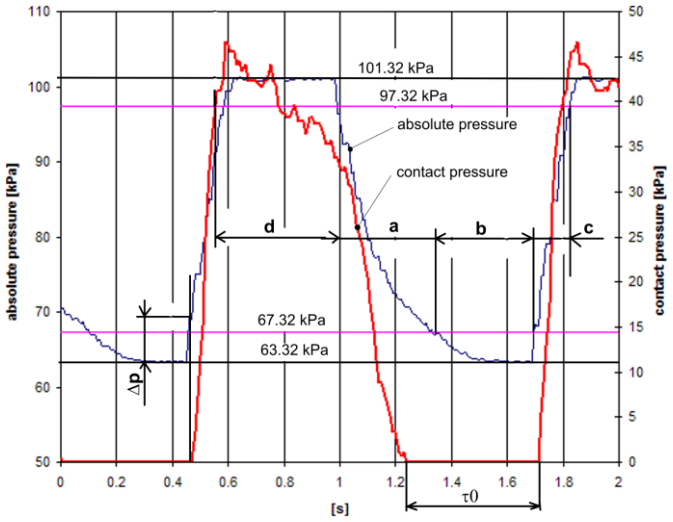

Figure 6. Intermittent vacuum and contact pressure chart

3.1. Pulsation rate and ratio

The results concerning the pulsation rate and ratio are presented in Table 1; the values were obtained using the averaged values for the durations of the pulsation phases, with the formulae (1) and (2). According to the recorded data, differences were recorded between the theoretical (set) pulsation rates and the achieved (real) ones (which were lower), but, for most of the operating conditions (eleven of the twelve tested), the relative differences did not exceed 5% (for the rubber liner, at 50 cycles/min and 60/40 pulsation ratio, the recorded pulsation rate was with 5.96% lower than the set rate). As far as the pulsation ratio was concerned, the relative differences between the values prescribed by the computer software (pulsator ratios) and the recorded ones (pulsation ratios) were lower than 3% for ten of the twelve variants.

Table 1. Results concerning the pulsation ratio and rate (average values)

Set pulsation ratio [%]	60			50		
Set pulsation rate [cycles/min]	50	55	60	50	55	60
WestfaliaSurge Classic Pro silicone liner						
Measured ratio [%]	60.6	60.1	59.0	50.0	50.2	50.2
Measured rate [cycles/min]	47.7	53.6	57.2	48.6	52.8	58.3
WestfaliaSurge Classic rubber liner						
Measured ratio [%]	56.1	59.0	58.7	48.2	48.8	49.4
Measured rate [cycles/min]	47.02	53.1	57.7	48.0	52.3	58.4

3.2. Duration of the cycle phases

The results referring to the duration of the cycle phases are summarized in Table 2. The pulsation rate and ratio had little effect over the duration of the a phase (an average duration of 0.348±0.005 s), but higher pulsation ratios led to longer b phases and shorter d phases. Increasing the pulsation rate led to shorter b and d phases. Some differences between the two types of teatcup liners were recorded for the b and d phases: the Classic Pro liner led to longer b phases and shorter d phases compared to the Classic liner (Figures 7 and 8).

Table 2. Results concerning the duration of the pulsation phases

Pulsation ratio [%]	60			50		
Pulsation rate [cycles/min]	50	55	60	50	55	60
WestfaliaSurge Classic Pro silicone liner						
Phase a [s]	0.353	0.340	0.333	0.337	0.347	0.330
Phase b [s]	0.410	0.333	0.287	0.280	0.223	0.187
Phase c [s]	0.123	0.107	0.120	0.117	0.113	0.113
Phase d [s]	0.373	0.340	0.31	0.5	0.453	0.400
WestfaliaSurge Classic rubber liner						
Phase a [s]	0.353	0.350	0.340	0.350	0.343	0.340
Phase b [s]	0.363	0.307	0.270	0.253	0.217	0.167
Phase c [s]	0.137	0.100	0.113	0.110	0.127	0.117
Phase d [s]	0.423	0.353	0.317	0.537	0.460	0.403

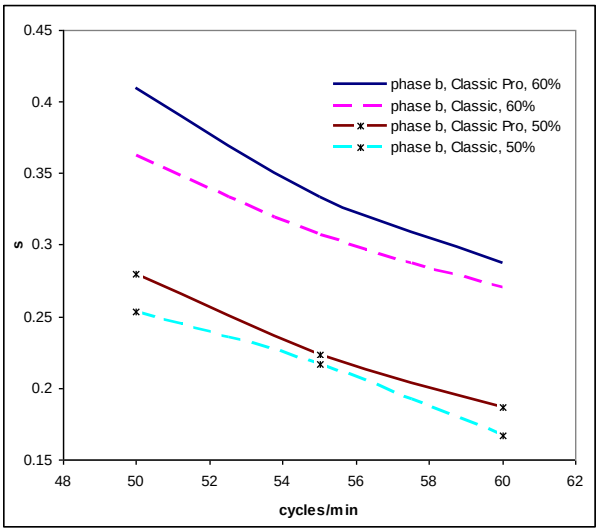

Figure 7. Duration of the b phase

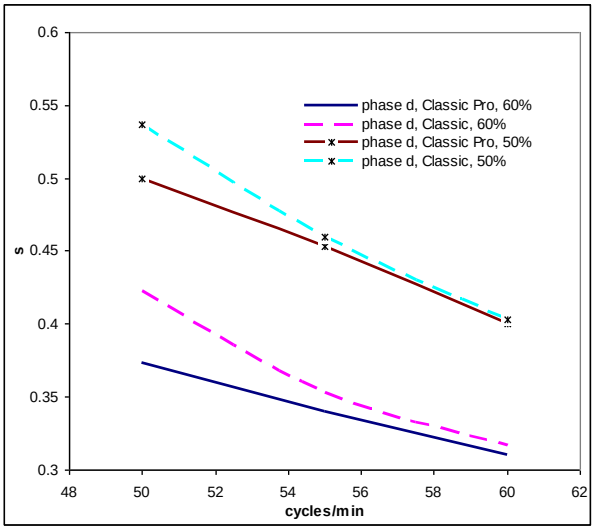

Figure 8. Duration of the d phase

The duration of the c phase was practically unaffected by the pulsation rate and ratio or by the type of teatcup liner; the average duration of the c phase was 0.1164±0.0029 s. This result is in

accordance with the results obtained in an experiment conducted by Bade, Reinemann, Zucali, Ruegg & Thompson (2009), in which 77 Holstein cows were investigated; the conclusion was that the duration of the c phase was not affected when some of the working parameters were modified.

3.3. The maximum liner-teat contact pressure

The results concerning the maximum liner-teat contact pressure are displayed in Figure 9 (standard error bars are also shown). The Classic Pro teatcup liner achieved much lower contact pressures compared to the Classic liner. According to van der Toll, Schrader & Aernouts (2010), the lower liner tension of the Classic Pro liner (5.4 % stretch, compared to the 6.8 % stretch of the Classic liner) was one of the causes that led to lower maximum contact pressures; Davis, Reinemann & Mein (2001) also noticed that the compressive load over the teat increased with liner tension. The thicker wall and the softer material of the silicone liner were the other factors leading to the reduction of the compressive load over the teat (Muthukumarappan & Reinemann, 1993).

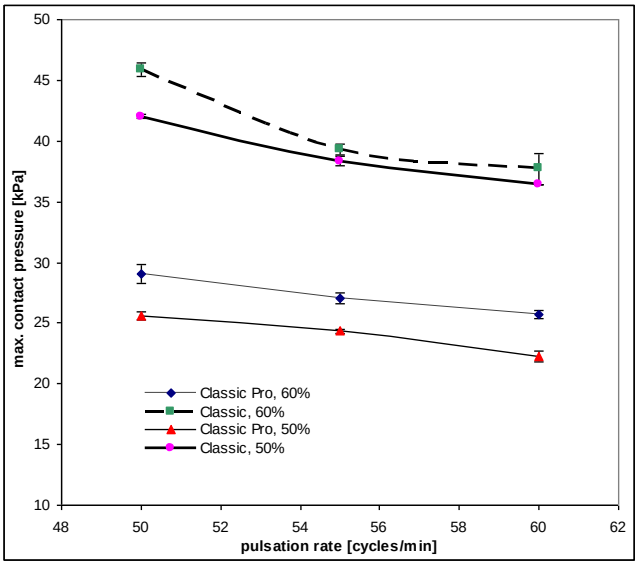

Figure 9. Maximum teat-liner contact pressure

For the both types of teatcup liners, higher contact pressures were recorded for the 60% pulsation ratio than for the 50% pulsation ratio and the maximum contact pressure decreased when the pulsation rate was increased; this tendency was more clear for the Classic rubber liner, lower contact pressures being recorded at higher pulsation rates. In the case of the Classic Pro silicone liner, the pulsation rate had a less important effect over the maximum contact pressure.

3.4. The critical collapsing pressure difference

It was found that the collapsing pressure difference was not significantly affected by the pulsation rate and ratio, but was influenced by the type of liner; the average values were 10.06±0.12 kPa for the Classic rubber liner and 14.21±0.47 kPa for the Classic Pro silicone liner. It should be mentioned that, in a "static" test (no pulsation, vacuum applied inside the liner), the pressure difference needed to close the liners was 12.7 kPa for the rubber one and 18.6 kPa for the silicone liner. This result is probably due to the different elastic properties and the thicker wall of the silicone liner, because, should only the mounting tension be considered, the lower mounting tension of silicone liner (5.4 % elongation, compared to 6.8 % for the Classic liner) would normally result in lower pressure differences (Mein, Williams & Thiel, 1987).

The higher critical collapsing pressure difference (of the Classic Pro liner) implies a higher touch point TP - the pressure difference required to collapse the liner to the point where the opposing walls touch each other (Mein& Reineman, 2009) - and a higher touch point leads to a lower compressive load over the teat (Spencer, Shin, Rogers & Cooper, 2007), which is in agreement with the results regarding the maximum teat-liner contact pressure (Figure 9).

It should be noted that the recorded values of the critical collapsing pressure difference were within the range of values reported by other authors - 11.9 kPa pressure difference according to Spencer & Jones (2000).

3.5. The time the teatcup liner is completely open and respectively closed

As mentioned before, the time while the teatcup liner is completely open was considered to be the time during which no force was applied to the teat; the results concerning this item are summarized in Table 3, together with the durations of the b phase of the pulsation cycle.

Table 3. Results concerning the time the teatcup liner is completely open

Pulsation ratio [%]	60			50		
Pulsation rate [cycles/min]	50	55	60	50	55	60
WestfaliaSurge Classic Pro silicone liner						
Phase b [s]	0.410	0.333	0.287	0.280	0.223	0.187
Time the liner is open [s]	0.520	0.417	0.390	0.363	0.320	0.273
WestfaliaSurge Classic rubber liner						
Phase b [s]	0.363	0.307	0.270	0.253	0.217	0.167
Time the liner is open [s]	0.477	0.370	0.363	0.377	0.330	0.263

As expected, increased pulsation rates and decreased pulsation ratios led to shorter periods of time during which the liner was completely open.

For both liners the time the liner was completely open was longer than the duration of the b phase of the pulsation cycle, for all the operating conditions; this means that the liner was already opened before the end of the a phase and remained opened at the beginning of the c phase.

At 60% pulsation ratio, the silicone liner was completely open for a longer time than the rubber one; at 50% pulsation ratio, the rubber liner was opened for a longer period. This results are correlated with the ones regarding the duration of the b phase (Figure 7): at 60/40 pulsation ratio, the b phase for the silicone liner was significantly longer than for the rubber liner, and, as a consequence, the Classic Pro liner was completely open for a longer time; at the 50/50 pulsation ratio, lower differences were recorded between the two liners in the terms of the duration of the b phase.

Figure 10 presents the duration of the c+d phases, for the both liners. The duration of the c+d phases decreased when the pulsation ratio and rate were increased, as expected.

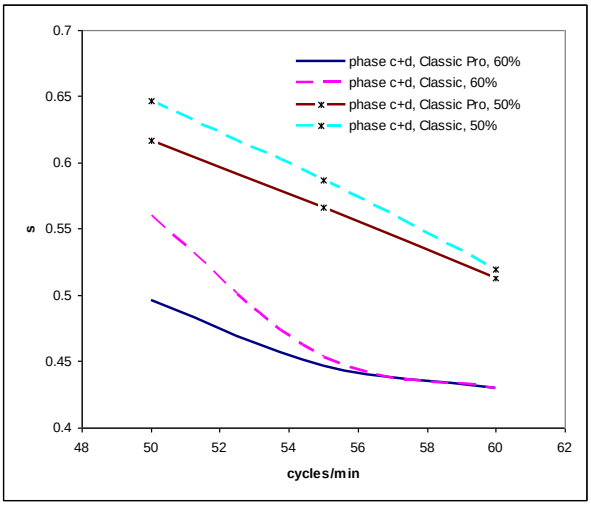

Figure 10. **Duration of the c+d phases**

For all the tested variants, the Classic liner was closed for a longer period than the Classic Pro liner, which is in accordance with the results recorded for the duration of the d phase (Figure 8); taking into account the findings of Adley & Butler (1994), this behavior might have also contributed to the higher liner-teat contact pressures (see Figure 9).

4. Conclusions

A computer controlled system for the evaluation of the mechanical milking machines; the system contains a computer driven pulses generator and a computer controlled pressure recording system. The system is a relatively low cost one and can be used for real life testing of milking machines.

Two types of teatcups were tested in order to evaluate the performances of the system, at different pulsation rates and ratios. The experimental results showed that for a prescribed pulsation ratio, the achieved pulsation ratio could be considered constant, as only minor variations were recorded. The recorded pulsation rate was also constant, although the achieved values were lower than the set ones. These results confirm the functionality of the computer driven pulses generator.

The results concerning the maximum liner-teat contact pressure showed that lower contact pressures were achieved by the silicone liner compared to the rubber one, due to the different properties of the material; the results concerning the critical collapsing pressure difference and the time the teatcup liner is completely open and respectively closed were consistent with the findings of other researchers, thus confirming the functionality of the pressure recording system.

References

Adley, NJD, & Butler, C 1994 ’Evaluation of the use of an artificial teat to measure the forces applied by a milking machine teatcup liner. *Journal of Dairy Research*, vol. 61, no 4, pp. 467-472. doi: 10.1017/S0022029900028399

ASAE EP445.1 1996, *Test equipment and its application for measuring milking machine operating characteristics*. American Society of Agricultural Engineers, St. Joseph, MI, USA

Bade, RD, Reinemann, DJ, Zucali, M, Ruegg, PL, & Thompson, PD 2009 ’Interactions of vacuum, b-phase duration and liner compression on milk flow rates in dairy cows. *Journal of Dairy Science*, vol. 92, no. 3, pp. 913-921. doi: 10.3168/jds.2008-1180

Billon, P, & Gaudin, V 2001 ’Influence of the duration of the a and c phases of pulsation on the milking characteristics and on udder health of dairy cows. *ICAR Technical Series*, no. 7, pp. 105-111.

Davis, MA, Reinemann, DJ, & Mein, GA 2001 ’Development and testing of a device to measure the compressive teat load applied to a bovine teat by the closed teatcup liner. *ASAE Paper Number 013007*, presented at the 2001 ASAE Annual International Meeting, Sacramento, California. Available from: www.uwex.edu/uwmril/pdf/MilkMachine/Liners/07 NMC CL Variation[1].pdf [20 February 2015].

doi: 10.13031/2013.7407.

Demba, S, Elsholz, S, Ammon C, & Rose-Meierhöfer, S 2016 'The usability of a pressure-indicating film to measure the teat load caused by a collapsing liner. *Sensors*, vol. 16, no. 10, 1597. doi:10.3390/s16101597

Gates, RS, & Scott, NR 1986 'Measurements of effective teat load during machine milking. *Transactions of the ASAE*, vol. 29, no. 4, pp. 1124-1130.

ISO 3918:2007, *Milking machine installations – Vocabulary*, International Organization for Standardization, Geneva, Switzerland.

ISO 5707:2007, *Milking machine installations - Construction and performance*, International Organization for Standardization, Geneva, Switzerland.

ISO 6690:2006, *Milking machine installations – Mechanical tests*, International Organization for Standardization, Geneva, Switzerland.

Kochman AK, Laney C, & Spencer SB 2008 'Effect of the duration of the c phase of pulsation on milking performance. Presented at the 47th National Mastitis Council Conference. Available from: http://www.laurenagrisystems.com/PDF/Research/EffectsOfDurationOfCPhaseOfPulsation.pdf [10 September 2014].

Mein, GA, Williams, DM, & Thiel, CC 1987 'Compressive load applied by the teatcup liner to the bovine teat. *Journal of Dairy Research*, vol. 54, no. 3, pp. 327-337. doi: 10.1017/S0022029900025504

Mein GA, & Reineman, DJ 2009' Biomechanics of Milking: Teat-Liner Interactions. *ASABE Paper Number 09743*, presented at the 2009 ASABE Annual International Meeting, St. Joseph, Michigan. Available from: www.uwex.edu/uwmril/pdf/MilkMachine/Liners/09 ASABE Mein-Reinemann teat Liner.pdf [10 September 2014]. doi: 10.13031/2013.27436

Muthukumarappan, K, & Reinemann, DJ 1993 'Compressive load applied by the teatcup liner to the bovine teat. *ASAE Paper no. 933538* presented at the 1993 International Winter Meeting of ASAE, Chicago, Illinois, ASAE, 2950 Niles Rd., St. Joseph, Michigan.

Reinemann, DJ, Mein, GA, & Muthukumarappan, K 1994 'Forces applied to the bovine teat by the teatcup liner during machine milking. *Report n. 94-D-052*, presented at the XII CIGR World Congress and AGENG'94 Conference on agricultural engineering, Milano, Italy. Available from: www.uwex.edu/uwmril/pdf/milkmachne/liners/94cigr_teat_load.pdf [10 September 2014].

Reinemann, DJ, Rasmussen, MD, & Mein, GA 2001 'Instrument requirements and methods for measuring vacuum in milking machines. *Transactions of ASAE*, vol. 44, no. 4, pp. 975-981.

Spencer, SB, & Jones, RL 2000 'Liner wall movement and vacuum measured by data aquision. *Journal of Dairy Science*, vol. 83, no. 5, pp. 1110-1114. doi: 10.3168/jds.S0022-0302(00)74976-9

Spencer, SB, Shin, J-W, Rogers, GW, & Cooper, JB 2007 'Effect of vacuum and ratio on the performance of a monobloc silicone milking liner. *Journal of Dairy Science*, vol. 90, no. 4, pp. 1725-1728. doi: 10.3168/jds.2006-493

van der Toll, PPJ, Schrader, W, & Aernouts, B 2010 'Pressure distribution at the teat-liner and teat-calf interfaces. *Journal of Dairy Science*, vol. 93, no. 1, pp. 45-52. doi: 10.3168/jds.2008-1864

Hungarian Association of Agricultural Informatics
European Federation for Information Technology in Agriculture, Food and the Environment

Journal of Agricultural Informatics. Vol. 8, No. 1
journal.magisz.org

A study on the digitization of supply chains in agriculture - an Indian experience

Suresh K Mudda[1], Chitti B Giddi[2], Murthy PVGK[3]

I N F O
Received 20 Apr. 2016
Accepted 20 Aug. 2016
Available on-line 15 Mar. 2017
Responsible Editor: M. Herdon

Keywords:
India, Supply chain, Digitization, ICT

A B S T R A C T

In the present day context of globalisation, changing information needs of the farmers, increasing pressure of population on the food security system, encouraging the developing economy like India to look for various alternatives in supply chain management and its digitization for its efficient and sustainable agricultural development. India is likely to be considered as the food basket to the world constituting 52% of total land under cultivation as compared to global average of 11%. It is also producing 134.5 MT of fruits and vegetables but due to inadequate cold storage and preservation facilities and improper supply chain infrastructure; there is enormous loss of wastages. Supply chains are principally concerned with the flow of products and information between supply chain member organizations procurement of materials, transformation of materials into finished products, and distribution of those products to end customers. Today's information-driven, integrated supply chains are enabling organizations to reduce inventory and costs, add product value, extend resources, accelerate time to market, and retain customers. Information Technology has started its dent in certain rural livelihoods especially the farmers in developing countries like India. IT can also do wonders in empowering small and marginal farmers who are operating in a complex, diverse and risk prone environment, who have poor access to information, especially regarding the production systems, customers and markets. In India, the limiting factors for farmers wanting to maximize their farm incomes are poor market linkages, poor access to quality farm-inputs, services and technology, lack of information about Government resources, institutions and extension services. ICT systems have pivotal role to play in market led extension activities. ICT s can connect the producers with buyers to initiate and sustain long term, mutually beneficial and sustainable professional relationships.

1. Introduction

In the present day context of globalisation, changing information needs of the farmers, increasing pressure of population on the food security system, encouraging the developing economy like India to look for various alternatives in supply chain management and its digitization for its efficient and sustainable agricultural development. Information Technology has started its dent in certain rural livelihoods especially the farmers in developing countries where India has no exception. It can also do wonders in empowering small and marginal farmers who are operating in a complex, diverse and risk prone environment, who have poor access to information, especially regarding the production systems, customers and markets. In India, the limiting factors for farmers wanting to maximize their farm incomes are poor market linkages, poor access to quality farm-inputs, services and technology, lack of information about Government resources, institutions and extension services. Internet is a faster and less expensive ever increasing speed mode of communication frequently used in IT for remote rainforest villages as compared to traditional communication services, such as mail and telephones. E-centres can help to improve social and economic opportunities in isolated areas, facilitate communication between indigenous peoples and organizations, and raise awareness of their concerns

[1] Suresh K Mudda
Acharya N G Ranga Agricultural University, Hyderabad, India.
skmudda@rediffmail.com
2 Chitti B Giddi
INGRAIN, Hyderabad, India.
3 Murthy PVGK
Acharya N G Ranga Agricultural University, Hyderabad, India

to mainstream society. In Asian countries, information technology and telecommunications have assumed an ever-increasing role in the creation of wealth at all levels.

The farmers also lack real time information about consumers, market demand and prices and hence are prone to more exploitation by existing intermediaries in the supply chain. With the growth of organized retailing and free global trade, farming is becoming highly knowledge intensive, commercialized, competitive and globalised, making it necessary to digitize, rebuild competitive and efficient agri-supply chains to benefit both the farmer as well as the consumer.

Digitization has a pivotal role to play in market led extension activities. ICTs can connect the producers with buyers to initiate and sustain long term, mutually beneficial and sustainable professional relationships. The existing disconnection between the producers and buyers in terms of harmonization of standards of agricultural produce is the cause for low value realization for the producers. Practically, the same disconnection is also increasing the cost of procurement for the buyers of agricultural commodities. Digitization helps to integrate the production, post-harvest management, value addition and marketing of agricultural produce. The ICT sphere also encompasses the quality aspects, agronomic aspects, traceability aspects and details of measurement of active ingredients, nutrition values etc. In the marketing and value addition perspective, usage of ICT to digitize the information contributes to increased efficiency and value enhancement of agricultural supply chains. Specially, in the Indian context, where the number of land holdings is small and big in number, reaching the multitudes of small farmers is the key for future food security. In a way the disintermediation in the supply chains can be possible by application of ICTs in adequate levels. The supporting mechanism or social benefits that are aimed to reach farmers can also be effectively implemented with the help of ICTs. In this background an effort has been made to study the digitization of supply chain management for its effectiveness in Indian context.

1.1. Indian agriculture

India is likely to be considered as the food basket to the world constituting 52% of total land under cultivation as compared to global average of 11%. It is also producing 134.5 M T of fruits and vegetables but due to inadequate cold storage and preservation facilities and improper supply chain infrastructure, there is enormous loss of wastages. Agriculture and its allied industries sector employs 67% of the country's population. In the post WTO regime, an effective agricultural marketing system through cost effective supply chain management, is the key driver of the sustainable development of agricultural economy. Agriculture has been the backbone of Indian economy since independence and before that, right now with nearly 12 per cent of the world’s arable land, India is the world’s third-largest producer of food grains, the second- largest producer of fruits and vegetables and the largest producer of milk; it also has the largest number of livestock. Add to that a range of agro- climatic regions and agri-produce, extremely industrious farmers, a country that is fundamentally strong in science and technology and an economy which one of the largest in the world with one of the highest growth rate and you should have the makings of a very good harvest. Yet the comprehensive outlook for Indian agriculture is far more complex than those statistics might suggest. Having just extricated itself from a period of negative growth of -0.1 percent in 2008-2009, to rise to an unspectacular 0.4 per cent in 2009-2010 with upward revision in the production, ‘agriculture, forestry and fishing’ sector in 2010-11 has shown a growth rate of 6.6 per cent, as against the growth rate of 5.4 per cent in the Advance estimates. Adjusted for inflation, even this 6.6 percent growth looks unexciting when compared to the growth rates in services and manufacturing. Today, agriculture accounts for 13.8 percent of the country’s gross domestic product, compared to 51 percent in the 1950s (Government of India, 2011). Worse, India is amongst the world’s largest wasters of food and faces a potential challenge to provide food security to its growing population in light of increasing global food prices and the declining rate of response of crops to added fertilizers. The reforms of 1991 have introduced Indian agriculture to the globalization which has very significant impact on agriculture and supply chain.

1.2. Supply chain management

Supply chains are principally concerned with the flow of products and information between supply chain member organizations - procurement of materials, transformation of materials into finished

products, and distribution of those products to end customers. Today's information-driven, integrated supply chains are enabling organizations to reduce inventory and costs, add product value, extend resources, accelerate time to market, and retain customers.

The real measure of supply chain success is how well activities coordinate across the supply chain to create value for consumers, while increasing the profitability of every link in the supply chain. In other words, supply chain management is the integrated process of producing value for the end user or ultimate consumer. The supply chains of different agricultural commodities in India, however, are fraught with challenges stemming from the inherent problems of the agriculture sector. The agri supply chain system of the country is determined by different sartorial issues like dominance of small/ marginal farmers, fragmented supply chains, absence of scale economies, low level of processing/value addition, inadequacy of marketing infrastructure etc. The agri supply chains in India and their management are now evolving to respond to the new marketing realities thrown by the wave of globalisation and other internal changes like rise in the level of disposable income of consumers, change in the food basket of the consumers towards high value products like fruits, vegetables and animal protein. The new challenges of the agricultural economy of the country have now spurred the government agencies to go in for different legal reforms for enabling and inviting private investment in agricultural marketing infrastructure, removing different entry barriers to promote coordinated supply chain and traceability.(Sazzad.P 2014).The amended APMR Act, the major agricultural Marketing Act of the country, being implemented by the different states of India, now contains enabling provisions to promote contract farming, direct marketing and setting up of private markets (hitherto banned). These measures will go a long way towards providing economies of scale to the small firms in establishing direct linkage between farmers, and processors/ exporters/ retailers, etc. Thus, the measure will provide both backward and forward linkages to evolve integrated supply chains for different agri produce in the country (MANAGE 2013).

Marketing channels for fruits and vegetables in India vary considerably by commodity and state, but they are generally very long and fragmented. The majority of domestic fruit and vegetable production is transacted through wholesale markets although depending on the state and commodity; farmers may sell to traders directly at the farm gate, to traders at village markets, or directly to processors, co-ops and others.

1.3. Coordinated supply chains

Coordinated supply chains involve structured relationships among producers, traders, processors, and buyers whereby detailed specifications are provided as to what and how much to produce, the time of delivery, quality and safety conditions, and price. These relationships often involve exchanges of information and sometimes assistance with technology and finance. Coordinated supply chains fit well with the logistical requirements of modern food markets, especially those for fresh and processed perishable foods. (Ahya, 2006). These chains can be used for process control of safety and quality and are more effective and efficient than control only at the end of the supply chain. Several companies in India are beginning to invest in integrated supply chain management systems and infrastructure with emphasis on quality and, to a lesser extent, on safety. Different models are emerging including fruit and vegetable retail outlets that directly procure produce from farmers or grower associations through various formal/informal contractual arrangements. Collection-cum-grading centres have been established in rural areas with all produce moving through a central distribution facility having modern infrastructure including cold storage, ripening rooms and controlled atmosphere chambers. Growers are required to follow certain specifications and are often provided with some inputs and technical advice about agronomic and post-harvest practices (MANAGE, 2013).

Contract farming for fruits and vegetables is already being practiced in several states and is likely to expand considerably due to legal reforms initiated in India, i.e., implementation of Model APMC Act. Until recently, contract farming was not legally recognized in most states and a legal framework for governing contracting arrangements was missing. Under the APMC Model Act a new chapter on 'contract farming' was added which provides for the registration of contract buyers, the recording of contract farming agreements and time-bound dispute resolution mechanisms. This information has

been digitized and kept ready as a blue print for further reference. It also provides an exemption from the levy of market fees for produce covered by contract farming agreements and provides indemnity to farmers' land to safeguard against the loss of land in the event of a dispute. Contract buyers will now be able to legally purchase commodities through individual purchase contracts or from farmers markets. Provision has also been made in the legislation for direct sale of farm produce to contract buyers from farmer's fields without it being routed through notified markets. This calls for the collective action in supply chains.

Initiatives are taken to establish more terminal markets based on modern infrastructure. Terminal markets would endeavour to integrate farm production with buyers by offering multiple choices to farmers for sale of produce such as electronic auctioning and facility for direct sale to exporter, processor and retail chain network under a single roof. In addition, the market would provide storage infrastructure thus offering the choice to trade at a future date to the participants. It is envisaged to offer a one-stop-solution that provides Logistics support including transport services & cool chain support and facility for storage (including warehouse, cold storage, ripening chamber, storage shed), facility for cleaning, grading, sorting, packaging and palletisation of produce and extension support and advisory to farmers.

The model presents integration of agri supply chains for perishables through terminal markets. Presently in the regime of fragmented and inefficient agri supply chains there is no control and command of chain partners on the other following that they are not able to maintain quality of produce in their chain. In order to bring integrated command, source quality produce by way of organizing farmers in groups and providing them the right technical advice and link farmers to the market, modern terminal market complexes will prove a dent. With increasing private investment in the food retail sector and impending changes in contract and marketing laws, shorter and more direct supply chains with traceability are expected to become more common. The incidence and spread of coordinated supply chains will be closely connected with the pace and direction of food retail sector modernization within India. Thus far, changes in food retail have been gradual, and considerably slower than observed in many other developing countries. Supermarket procurement regimes for sourcing of fruits, vegetables, dairy and meat strongly influence the organization of the supply chains. The rising scale of organized retail in the Asian countries (like Metro Cash & Carry, Tata Chemicals and Field Fresh Foods, Bharti Enterprises, Reliance Fresh in India) is now playing a vital role in organizing farmer production bases and integrating these into the retailers' fresh produce supply chain, thus procurement systems in this segment is changing fast responding to the consumer demand and competition.

2. Digitization and its implications

ICT is a powerful tool to integrate the production, post-harvest management, value addition and marketing of agricultural produce. The ICT sphere also encompasses the quality aspects, agronomic aspects, traceability aspects and details of measurement of active ingredients, nutrition values etc. In the marketing and value addition perspective, ICT contributes to increased efficiency and value enhancement of agricultural supply chains. Specially, in the Indian context, where the number of land holdings is small and big in number, reaching the multitudes of small farmers is the key for future food security. In a way the disintermediation in the supply chains can be possible by application of ICTs in adequate levels. These IT applications cannot be limited to marketing aspects alone but are to be integrated with the production aspects for its sustainable development. These aspects are having certain social implications in Indian context. The existing land use patterns, land records, tenancy norms soil health and its enrichment are being digitized and also need to be recorded. Information Technology should be used for maintaining an updated and enriched database of region specific agricultural information and timely dissemination of the information pertaining to seed selection, actions relating to arrival of monsoon, climate control etc. to the farmers. In addition, information regarding agricultural products, demand-supply status in respect of different products and the current price should be made available on-line to the farmers for taking timely decisions on crop product diversification strategies and positioning of the same in right market to get optimum revenue. With

agile, demand-driven supply, focusing on reducing end-to-end supply network time by building a flexible and responsive supply network is the need of the hour. (Narula, et al 2010)

The educational and professional institutions should take for guiding the latest information using IT as a tool and make it available to the farmers. The need of the day is to harness the vast potential of agriculture in Indian economy.(ICT source book)

The ICT sphere also encompasses the quality aspects, agronomic aspects, traceability aspects and details of measurement of active ingredients, nutrition values etc. In the marketing and value addition perspective, ICT contributes to increased efficiency and value enhancement of agricultural supply chains. Specially, in the Indian context, where the number of land holdings is small and big in number, reaching the multitudes of small farmers is the key for future food security. In a way the disintermediation in the supply chains can be possible by application of ICTs in adequate levels. The supporting mechanism or social benefits that are aimed to reach farmers can also be effectively implemented with the help of ICTs. In conditions of poor information flows supply chains are highly fragmented. Otherwise information technology driven innovations make it easier to acquire, manage, and process information and allow closer integration between adjacent steps in the value chain. There is therefore greater integration of supply chains based on information availability (Kunaka, 2010)

Information on supply chain management is a basic element in any development activity. Once it is digitized Information will be available and accessible to all, be it scientific, technical, economic, social, institutional, administrative, legal, historical or cultural in nature. Agricultural information is useful only if it is available, if the users have access to it, in the appropriate form and language. What do the farmers want? They require information *inter alia* on supply of inputs, new technologies, early warning systems (drought, pests, diseases), credit, market prices etc. Information, in the field of agriculture, to be of benefit, has to be tailored to local agro-ecological and socio-economic conditions. It has to be backed up by relevant input supply services, and public policies. Synchronization in time and space between knowledge and input delivery systems is essential to impart credibility to the extension message. Agricultural Extension Services do play an important role in delivering information, knowledge and advice to farmers. However to remain relevant in these changing times, it has to specialise in "effectively managing and transferring knowledge or information packages". Emerging digital technologies can play an important role in supporting extension in this regard. (Meera 2014). Over the past two decades, Governments all around the world have invested heavily in strengthening the national ICT infrastructure.

Need for improved agricultural extension throughout the developing world has never been greater. Agricultural and rural development and hence, rural extension continue to be in transition in the developing world. These transitions are happening because of the forces that are driving the world agriculture today. The vulnerability of farming in the developing world is quite evident due to forces like climate change, changes in natural resources quality, lack of coping strategies at micro and macro levels of decision making, coupled with globalization, emerging market forces like commodity markets, sustainability constraints etc. The challenges can only be met from information intensive efforts in the extension systems (Shaik Meera, N *et.al* 2010). These information intensive extension efforts can be possible when extension systems embrace digital opportunities available with us today.

2.1. E-choupal experience of ITC

ITC Ltd implemented a project on electronic market place for the soybean farmers in the state of Madhya Pradesh. The project owes its success to the factors such as utilization of local leadership in the villages, a sustainable business model and collaboration between the local authorities and the corporate implementer. The technology embarked was easy to replicate and easily scalable, and it was customized according to the needs of the local farmers. The project has helped the farmers developing sustainable income levels, elimination of the middlemen, developing easy access to the market place and shared ownership of the project (Figure 1).

Figure 1. Demonstration for farmers

2.2. e-governance in fisheries

The Fisher Friend Mobile Application (FFMA) is a unique, single window solution for the holistic shore-to-shore needs of the fishing community, providing vulnerable fishermen immediate access to critical, near real-time knowledge and information services on weather, potential fishing zones, ocean state forecasts, and market related information. The application is an efficient and effective decision support tool for the fisher community to make informed decisions about their own personal safety and the safety of their boats, as well as make smart choices for fishing and marketing their catch. FFMA is developed on an android platform in partnership with Wireless Reach Qualcomm and Tata Consultancy Services and is currently available in English, Tamil, and Telugu (Figure 2).

Figure 2. The Fisher Friend Mobile App

2.3. ICT platform of EID Parry

East India Distilleries (EID) Parry has implemented the project "Parry's Corner" to help the farmers, provide them with value-added services, and improve their income levels and the productivity of their farms. The self-help groups in the vicinity are using the ICT platform for e-commerce. This has helped in the creation of social networks. Social networks facilitate the diffusion of ICT platforms. The major reason for the success of this project has been that the company has been in operation in that region for a long time. The high levels of trust existing between the company executives and farmers helped in the rapid diffusion of the utilization of the ICT platform for a variety of reasons. The technology selected was a low cost option and hence the overheads were not high for platform.

As FAO (2005) notes, the information system obviously remove critical barriers that have kept farmers from participating in the commercial supply chain. Farmers receive relevant and timely information regarding crop production, the company effectively communicates demand and quality requirements, and farmers can demand a fair price and be assured of a market. Further, agricultural

yields, access to finance, agricultural extension services, and time required to transact with EID Parry all have reportedly improved. These improvements have not been quantified, however (FAO, 2005).

2.4. Farmers portal of government of India

The Farmers' Portal of the Department of Agriculture & Cooperation is a platform for farmers to seek any information related to agriculture (Figure 3). Detailed information on farmers' insurance, agricultural storage, crops, extension activities, seeds, pesticides, farm machineries, etc. is provided. Details of fertilizers, market prices, package and practices, programmes, welfare schemes are also given. Block level details related to soil fertility, storage, insurance, training, etc. are available in an interactive map. Users can also download farm friendly handbook, scheme guidelines, etc.

Figure 3. The Farmers' Portal of the Department of Agriculture & Cooperation

2.4. Kisan call centre services

Kisan Call Centres (KCCs) (Figure 4) was launched by the Ministry of Agriculture took to harness the potential of ICT in agriculture. This initiative was aimed at answering farmer's queries on a telephone call in their own language / dialect. IFFCO Kisan Sanchar Limited (IKSL) was selected by the Department of Agriculture and Cooperation (DAC), Ministry of Agriculture (MoA), Government of India, to manage the KCC services. The services were re-launched on 1st May 2014 by IKSL. In this endeavour, IKSL had completely revamped the services and set up state of the art ICT infrastructure (Figure 5).

Figure 4. A Kisan Call Centre

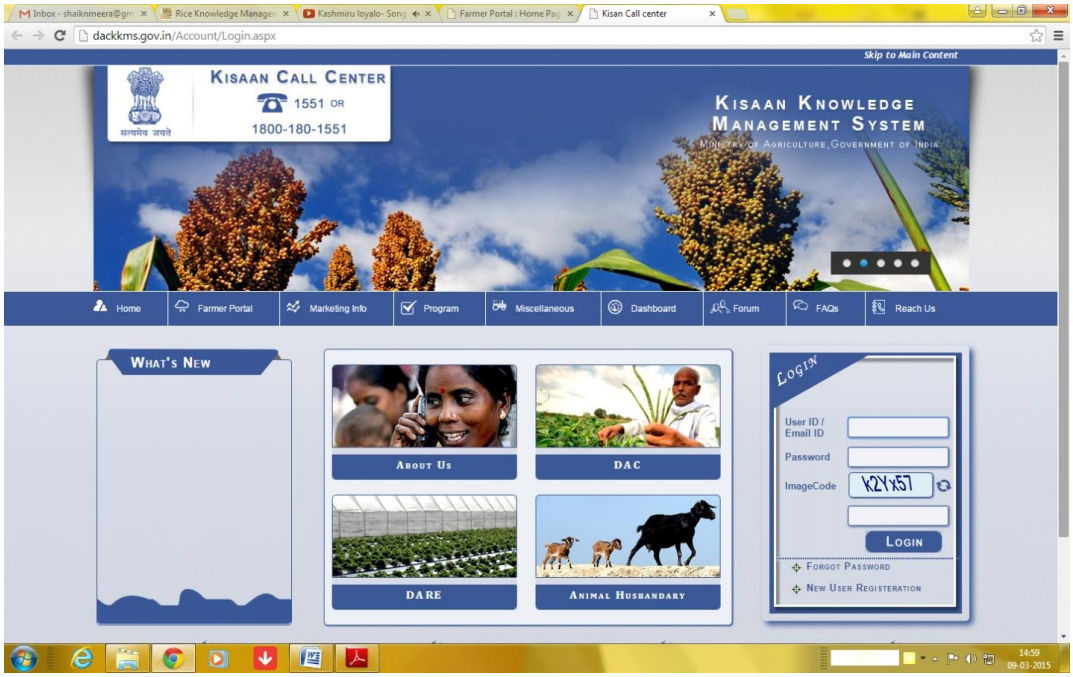

Figure 5. The Knowledge Management System

2.5. m-kisan Project of Government of India

The project conceptualized, designed and developed in-house within the Department of Agriculture & Cooperation USSD has widened the outreach of scientists, experts and Government officers posted down to the Block level to disseminate information, give advisories and to provide advisories to farmers through their mobile telephones. Since its inception nearly 72 crore messages or more than 210 crore SMSs have been sent to farmers throughout the length and breadth of the country. These figures are rising ever since.

These messages are specific to farmers' specific needs & relevance at a particular point of time and generate heavy inflow of calls in the Kisan Call Centres where people call up to get supplementary information. SMS Portal for Farmers has empowered all Central and State Government Organizations in Agriculture & Allied sectors (including State Agriculture Universities, KrishiVigyanKendras, Agromet Forecasts Units of India Meteorological Department, ICAR Institutes, Organization in Animal Husbandry, Dairying & Fisheries etc.) to give information/services/advisories to farmers by SMS in their language, preference of agricultural practices and locations.

USSD (Unstructured Supplementary Service Data), IVRS (Interactive Voice Response System) and Pull SMS are value added services which have enabled farmers and other stakeholders not only to receive broadcast messages but also to get web based services on their mobile without having internet. Semi-literate and illiterate farmers have also been targeted to be reached through voice messages.

3. Limitations

As the Supply Chain involves a number of players, the extent of integration of services depends on the degree of trust and information sharing amongst the players. It is often observed that the big players in their efforts to make vertical/horizontal integration of different activities end up gobbling up the weak ones. What in fact is called for is strengthening of the system and process, so that requisite synergies evolve to give benefits to all the partners.

The ultimate choice of the ICT enabled agriculture approach depends on (1) the ICT policy environment, (2) the capacity of ICT service providers, (3) the type of stakeholders and the ICT approach adopted, and (4) the nature of the local communities, including their ability to access and apply the knowledge and various e-readiness parameters. The level of integration of digital media into the governance process in agriculture will determine the fate of Indian agriculture in years to come.

In order to shore up the emergence of professionally managed agri-supply management of different agricultural produce, the Government should play its facilitating role.

Some of the major issues that need to be focused in the public domain are:

- Focus should be laid on free play of demand and supply forces in the market. This has to be enabled by removing different entry barriers, having a proper market information system, promoting grading and standardization, taking care of quality and safety issues, putting up a strong system of risk management and price formation mechanism. This can be done only by digitizing the information available at various levels of supply chain.
- Different legal restrictions inhibiting growth of competitive environment should be dismantled and replaced by a facilitating legal environment.
- Infrastructure is the major constraint in Indian marketing system. Since it is difficult to arrange sufficient funds from the public exchequer for the development of infrastructure facilities, the need of the hour is to explore different Public Private Partnership models.
- The extension mechanism of the country is production oriented relegating the marketing aspects to the backburners. It is time for the Stakeholders to provide basic information in supply chain in a digitized form.

Within broad framework of a conducive environment provided by Government side, the private sector should come up in a pro-active manner to invest in agriculture sector. In no way, they should be discouraged by the teething troubles as entrepreneurs in this virgin sector in India. The managerial efficiencies brought about by the private sector to the agricultural economy of the country will go a long way towards ensuring optimum utilisation of resources, thereby ensuring sustainable growth for the sector.

4. Conclusions

Usefulness of ICT in the form of digitising the all possible information is well established in improving productivity of Agricultural sector and this needs to be addressed by authorities. Food loss reduction is less costly than an equivalent increase in food production. If efforts are not made to modernize the harvest handling system for horticultural crops, then postharvest losses will continue to have a negative economic and environmental impact. There is no doubt that postharvest food loss reduction significantly increases food availability. An efficient collaboration between stakeholders will reduce risk, losses and greatly improve the efficiency to ensure food security and development.

The important link between the whole chain of digital networks and their applications is the ultimate beneficiaries of these initiatives those are the stakeholders. It is common to find that intended users (farmers) are either unaware of the ICT services / or do not perceive these services as applicable in their field conditions. Unfortunately the task of understanding the clientele and their information need has been subsided by the technological enthusiasm that is prevailing in Indian context. (Shalendra 2013)

Agricultural extension, whether public or private, operates in a context that influences the organization, form, and content of transfer activities. For instance, what necessitates current extension / advisory organisations to integrate digitization into their functional / structural components? The history and recent developments in Asia illustrate that ICT "prescriptions" are doomed to fail if they are not based on 'farmers needs'. And it must be driven by learning about what works and what does not and by the nature of local circumstances and context. We need to address relevant issues such as what makes public extension workers to become info-mediatory. Their job chart needs to be transformed radically with a scope for incentives for efficient performance using digital tools. ICT applications alone will not be readily available, accessible and applicable in farmers' conditions. It requires higher commitments from all the agricultural professionals. Further we need to build farmers communities on large scale government should plan campaigns for 'zooming in zooming out' farmers learning/ experiences using ICTs.

The idea of Digitization of supply chain, essentially provides linkages, enhance market access, improve business process, increase product diversity and reduce development cycle time in Indian agriculture. Understanding ICT context for Indian agriculture will help developing nation level strategies. Digital India's perspective of agriculture will have a real challenge in integrating 'knowledge' with 'time critical services' in the whole chain of agricultural value chain. We should have evidences of use, pattern, purpose, users etc., for ICT activities. These IT applications cannot be limited to marketing aspects alone but are to be integrated with the production aspects for its sustainable development. These aspects are having certain social implications in Indian context. The existing land use patterns, land records, tenancy norms soil health and its enrichment are being digitized and also need to be recorded. Information Technology should be used for maintaining an updated and enriched database of region specific agricultural information and timely dissemination of the information pertaining to seed selection, actions relating to arrival of monsoon, climate control etc. to the farmers. In addition, information regarding agricultural products, demand-supply status in respect of different products and the current price should be made available on-line to the farmers for taking timely decisions on crop product diversification strategies and positioning of the same in right market to get optimum revenue. With agile, demand-driven supply, focusing on reducing end-to-end supply network time by building a flexible and responsive supply network is the need of the hour.

The educational and professional institutions should take for guiding the latest information using IT as a tool and make it available to the farmers. The need of the day is to harness the vast potential of agriculture in Indian economy.

Table 1. Broken Links in Agri Supply Chain in India

Production	Supply Chain	Processing	Marketing
• Poor extension • Quality inputs • Low productivity • Deficient and inefficient production management • Non demand linked production • Improper post-harvest management resulting in poor quality	• Lack of storage • Poor transportation • High wastages • Multiple intermediaries • Fresh produce transported to mandis in open baskets or gunny bags stacked one on top of the other • Cold chain absent or broken, produce deteriorates rapidly • Food safety is major concern: Hygiene and pesticide MRL not monitored	• Low processing • Lack of quality • Poor returns • Low capacity utilization	• Poor Infrastructure • Lack of grading • No linkages • Non-transparency in prices • Long delays from producer to retailer
Each segment working in an isolated manner resulting in multiple losses across the value chain			

References

Ahya, C 2006 'The Retail Supply Chain Revolution'. The Economic Times, Retrieved on 29.03.2013 from http://articles.economictimes.indiatimes.com/2006-12- 07/news/27452233_1_retail-revolution-retail-sector-supply-chain.

FAO (Food and Agriculture Organization) 2005, 'Case Study: Community Based Information Systems', India.

Government of India 2011, Economic Survey report, Ministry of Finance.

http://www.manage.gov.in/studymaterial/scm/E.pdf .Reading material on Supply chain management in agriculture.

ICT applications for Agribusiness supply chains. http://www.ictinagriculture.org/sourcebook/module-10-ict-applications-agribussiness-supply-chains.

Kunaka, C 2011, 'Logistics in Lagging Regions: Overcoming Local Barriers to Global Connectivity', Washington DC, World Bank.

MANAGE, 2013, Reading material on supply chain management, Rajendranagar, Hyderabad, India.

Sapna, A, Narula, Navin Nainwal, 2010, ICT and agriculture supply chains, opportunities and strategies for successful implementation. Information technology in developing countries. A newsletter of the IFIP working group. 9.4 vol. 20, no.1. www.iimahd.ernet.in/e gov/ifip/feb2010/sapna.narula.htm.

Sazzad, P 2014, 'Food supply chain management in Indian Agriculture: Issues, opportunities and further research'. *African journal of management*, vol. 8, no.14, pp. 572-581. doi. 10.5897/ajbm2013.7292

Shaik Meera, N 2014, 'Digital India & Agricultural Sector - Can we make it Digital Bharat?' Concept paper, DRR, Rajendranagar, Hyderabad.

Shaik Meera, N, Arunkumar, S, Amtul Waris, Vara Prasad, C, Muthuraman, P, Mangalsen, and Vikranth, BC 2010, 'E-Learning in extension systems- Emperical study in Agricultural Extension in India'. Indian Journal of Extension Education, vol. 46, no. 3&4, pp. 94-101.

Shalendra and Purushottam, 2013, 'User centric ICT model for supply chain of horticulture crops in India'. *Agricultural Economics Research Review*, vol. 26, no. 1, pp. 91-100.

Hungarian Association of Agricultural Informatics
European Federation for Information Technology in Agriculture, Food and the Environment

Journal of Agricultural Informatics. Vol. 8, No. 1
journal.magisz.org

Evaluation of cellulose content by infrared spectroscopy

János Jóvér[1], Attila Nagy[2], János Tamás[3],

I N F O
Received 6 Feb.. 2017
Accepted 10 Mar. 2017
Available on-line 15 Mar. 2017
Responsible Editor: M. Herdon

Keywords:
Cellulose content, spectral curve, non-invasive measurements, Cellulose Absorption Index, applied informatics

A B S T R A C T

Cellulose is one of the most abundant organic chemical material in the world which is a raw material for several fields. In many cases this material is appearing as residue in the agriculture and forestry. Thus the application of cellulose as secondary raw material is desirable which requires a rapid estimation method reflecting the potential amount of cellulose stocks. The remote sensing based Cellulose Absorption Index (CAI) is an adequate method to make objective estimations of the quantity plant litters on soil surfaces. By the application of infrared spectroscopy CAI can be calculated to make rapid estimation of cellulose content in laboratory scale. In this research work cellulose contents of sweet sorghum bagasse were measured by the method of Van Soest. Based on the absorbance values determined in A.U. CAI values of the sweet sorghum bagasse samples were calculated. As a result a notable correlation (R^2=0.733) was found between the measured cellulose values and the modified Cellulose Absorption Index.

1. Introduction

Plant fibers have high strength, low density and high sustainability (Hepworth et al. 2000; Madsen & Lilholt 2003). According to Klinke et al. (2001) the fiber tensile strength and elastic modulus depend on the cellulose content squared. Cellulose is a linear structured glucose polymer which is the most abundant material appearing in the world (Granström 2009). As a chemical raw material, cellulose and its derivatives are widely utilized in many fields, such as the production of paper, textile and pharmaceuticals (Lavanya et al., 2011).

Primer biomass is widely used for energy purposes but the utilization of the secondary biomass sources can be more adequate in the sustainable energy production. Cellulose as a by-product of the agricultural industries or energy crops, is digestible by microorganisms for energy (Watanabe 2013) hence bioethanol produced from cellulosic biomass is a promising renewable energy source (Wagner & Kaltschmitt 2013). In contrast to first generation bioethanol, which is derived from sugar or starch, cellulosic ethanol may be produced from agricultural residues. The ability to make fuel from locally sourced, secondary non-food feedstocks makes cellulosic bioethanol an effective method for reducing dependence on fossil fuels.

Based on the utilization possibilities the determination of cellulose content in plants can be very important. Cellulose is a basic compound of the cell wall and appearing with hemicelluloses and lignin which make the determination process difficult. The determination of these fibers is based on chemical processes (Van Soest et al. 1991; Zhao et al. 2009; Erdei 2013; Khalil et al. 2015). Chemical determination of cellulose and other fibers of the cell wall is a difficult and time consuming process while in several fields of the agriculture the application of non-invasive spectral measurements provides rapid and reliable data (Riczu et al. 2012; Nagy et al., 2014). By the utilization of applied informatics spectral data can be visualized as spectral curves moreover several mathematical

[1] János Jóvér
University of Debrecen, 4032, Debrecen, Böszörményi út 138, Hungary
jover@agr.unideb.hu
[2] Attila Nagy
University of Debrecen, 4032, Debrecen, Böszörményi út 138, Hungary
attilanagy@agr.unideb.hu
[3] János Tamás
University of Debrecen, 4032, Debrecen, Böszörményi út 138, Hungary
tamas@agr.unideb.hu

opportunities are available. Thus the application of vegetation indices (e.g.: Leaf Area Index, Normalized Difference Vegetation Index, Enhanced Vegetation Index etc.) calculated by spectral data can be an adequate method to receive information rapidly about plant vegetation. In the field of cellulose content determination the Cellulose Absorption Index (CAI) can be utilized which was worked out to discriminate plant litter from soil based on the absorbance values of electromagnetic radiations in the range of 2000 nm to 2200 nm (Daughtry et al., 1996).

In this study the analysis of spectral curve of sweet sorghum bagasse was made so as to calculate a modified CAI. Index values were compared with the measured cellulose contents in order to evaluate reliability of the index values within laboratory scale. The goal of the current research work was to evaluate the application possibilities of CAI in the field of rapid determination of cellulose.

2. Materials and methods

Sorghum bagasse was derived from the Research Institute of Karcag. After chopping, the air-dry bagasse was grinded first with a rough-, then a fine grinder (Condux) in order to achieve the size from 0.1 mm to 2 mm. Fiber content determinations of the sorghum bagasse samples were conducted according to Georging & Van Soest (1975) where samples were analyzed to acid-detergent fiber fraction (ADF), neutral detergent fiber fraction (NDF) and acid detergent lignin (ADL) (Figure 1). Due to the time consuming characteristics of this analytical process (approximately 6,5 hours/sample) this preliminary research work was based on the results of 27 sorghum samples.

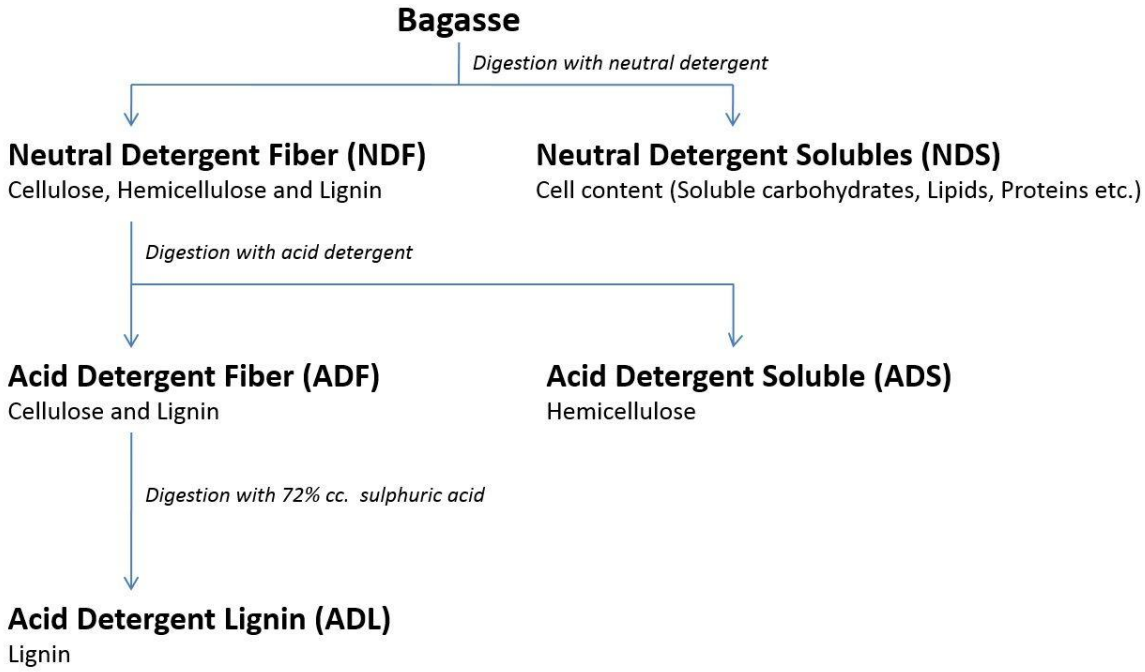

Figure 1. Determination scheme of carbohydrate fibers based on the findings of Van Soest

The calculation of cellulose content were done as follows:

$$\text{Cellulose content} = \text{ADF} - \text{ADL}$$

The absorbance measurements were carried out by AvaSpec NIR256-2.5-HSC Fiber Optic Spectrometer within 1000–2500 nm interval. During the measurement the average spectral resolution was 6.43 nm. The AvaSpec 2048 system consists of a spectrometer (detector) and connected by an 8 μm core diameter fiber optic standard AvaLight-HAL-MINI halogen light source (Figure 2). The accurate measurement was provided by a special spectral sampling black box, since the samples were

isolated from the external light. The end of the optical fiber was positioned 5 mm above the surface of each sample. Spectral data were managed by AvaSoft 8.2 software.

Figure 2. Light source, detector and sampling black box of the applied spectrometer

During the measurements the integration time was set as 57.03 ms while absorbance spectral data were worked out as averages of 30 measurements with a pixel smoothing number of 3. Spectral data were exported to Microsoft Office Excel in order to visualize spectral curves in a same plot and to make further calculations. Measurements were done on air-dry samples in order to avoid the alteration of spectral curves caused by water. Absorptions in the 2000 nm to 2200 nm range are sensitive to cellulose (Nagler et al. 2003), so that based on the measured spectral data within this interval the Cellulose Absorption Index was calculated (Eq. 1) which is a commonly used index to indicate exposed surfaces containing dried plant material (Daughtry 2001; Daughtry et al. 2004). This index provides the possibility to discriminate pure soil surfaces from pure scenes of some crop residues and tree litters in the case of arable lands (Daughtry et al. 1996; Nagler 1997). Based on the detected reflectance values this index ranges from -3 to more than 4. Positive values of CAI represent the presence of cellulose while a negative value means the absence of this chemical material.

Eq. 1 $$CAI = 0.5(\rho_{2000} + \rho_{2200})-\rho_{2100}$$

According to Elvidge (1988; 1990) 2100 nm, 2180-2220 nm and 2310-2380 nm spectral ranges are sensitive to lignocellulose compounds. Thus the presence of hemicellulose and lignin can modify the value of CAI.

In this research CAI values were calculated by the measured absorbance values determined in A.U. (Absorbance Unit) in order to evaluate the application possibilities of this index to determine cellulose content in plant materials if it is calculated by absorbance values directly. The relation of cellulose content and the modified CAI was evaluated by linear regression analysis using R statistical software with R Studio user surface (R Core Team 2016).

3. Results and discussion

Based on the measured data the average Acid Detergent Fiber content was 42.63% with the standard deviation of 4.32%. In the case of Acid Detergent Lignin the average value was 3.80±2.15%. According to the calculations cellulose contents were varied between 36.31% and 42.84%. The average cellulose content of the samples was 38.55% with a standard deviation value of 1.69% and 38.56% median value.

In the case of the spectral curves of the evaluated sweet sorghum samples two low points and a peak point were detected in the spectral bands which are used in the calculation of CAI. Low points were detected at the wavelength of 2000 nm and 2200 nm, while the peak point were detected at the wavelength of 2100 nm (Figure 3). The absorbance spectra of sweet sorghum samples with different cellulose content were alike regarding the shape, but differences were detected in the intensity of absorption.

Correlation between absorbance values detected on the wavelength's of 2000 nm (R^2=0.01), 2100 nm (R^2=0.04), 2200 nm (R^2=0.02) and cellulose contents were not found. The reason of this phenomenon can be the fact that CAI was worked out for remote sensing detecting soil covering plant litters. Thus CAI estimate the senescent plant covering of soils and do not assess directly the spectral features of cellulose content in plant materials. In this case the spectral features of soil surface stoutly prevail. Spectral characteristics of soils and plant materials are different therefore lands with different rate of plant litter covering can be distinguished based on their heterogenic spectral characteristics. In that case cellulose sensitivity is strongly corresponding with the plant litter ground cover, and not with the exact amount of cellulose content.

In this study the cellulose content of plant textures were evaluated, and cellulose contents were not heterogenic enough so as to detect differences among the evaluated samples in the spectral bands of CAI. Nevertheless further calculations were carried out by spectral bands of CAI based on the scientific results of Roberts et al. (1990, 1993), Daughtry et al. (1996) and Nagler et al. (2003).

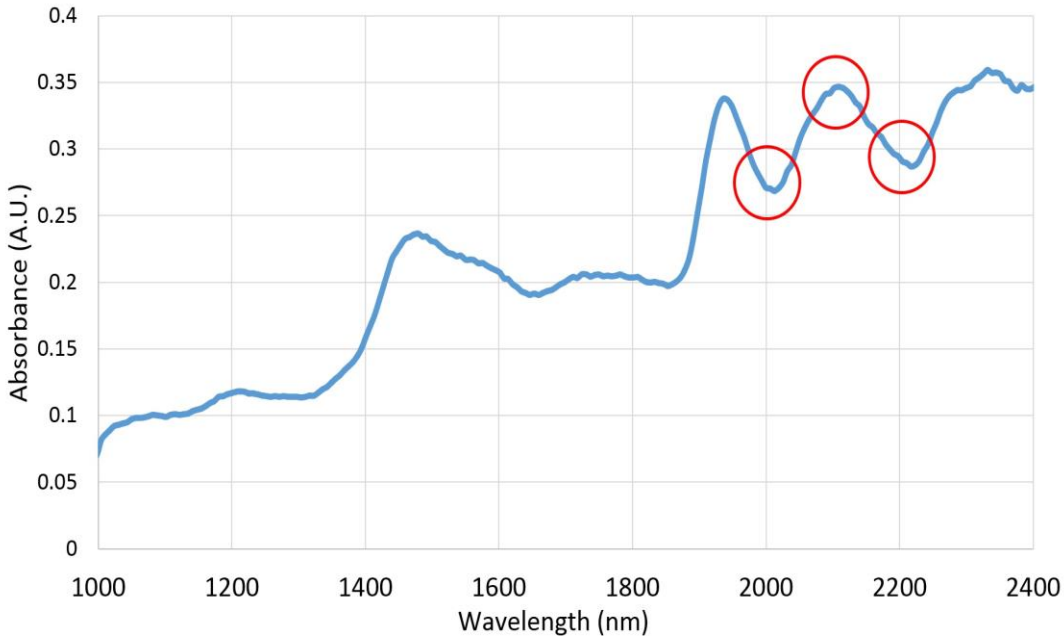

Figure 3. The spectral curve of the average absormance values

Regarding the average value of the calculated Cellulose Absorption Index value was -0.063 with the standard deviation of 0.009. CAI values between -0.078 and -0.042 were characteristic to the evaluated sweet sorghum bagasse samples. Negative CAI values basically referring to the absence of cellulose in so far as it is determined by reflectance. However this index is basically applied to quantify plant litter on soil surface the values showed a strong correlation with the calculated cellulose contents (R^2=0.733) nevertheless CAI values were negative (Figure 4). This phenomenon probably caused by the change of the reflectance values to absorbance values in the equation.

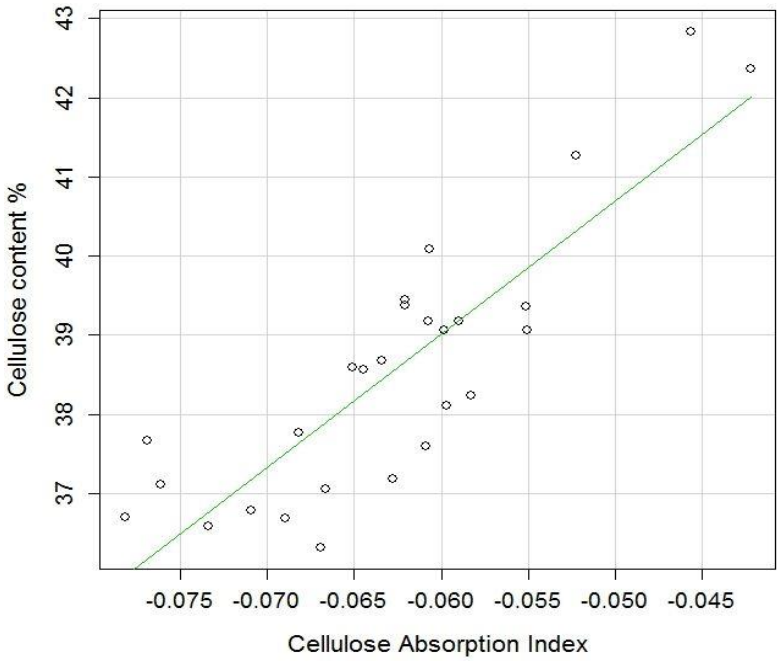

Figure 4. The correlation of Cellulose Absorption Index with the measured cellulose contents

In the study of Daughtry et al. (1996) CAI (calculated by reflectance) was applied to discriminate plant litters from soil surface and the absorption features of this index were described as the depth of the lignocellulose absorption feature. Nevertheless strict correlation with CAI was found only in the case of the cellulose content of these materials. Correlation of CAI values between lignin (R^2=0.01), hemicellulose (R^2=0.01) or total lignocellulose (R^2=0.05) contents were not found. Thus discrimination of plant litter from pure soil surface by CAI can be based on the cellulose content of these materials.

According to results Cellulose Absorption Index determined by absorbance can be adequate for the rapid estimation of cellulose content in plant materials so as to qualify plant materials. By this way of spectral analysis can result rapid result of cellulose content which can be a useful information in several specific agricultural fields.

3. Conclusions

Cellulose Absorption Index is a commonly used vegetation index in the field of remote sensing discriminating plant materials from pure surfaces. The determination of this index is based on the reflectance features of the spectral bands of 2000 nm, 2100 nm and 2200 nm. According to the results of this preliminary study Cellulose Absorption Index is an adequate value to estimate cellulose content in plant materials on laboratory scale. By the application of infrared spectroscopy with applied informatics tools rapid values can be received about cellulose contents. These results can useful in the process of plant breeding (e.g.: in breeding methods to eliminate risks caused by falling down), or in the field of bioenergy as well (Lee et al. 2017). According to some scientific results CAI is applicable to dry senescent plant materials but in the current study correlation was found only between CAI and cellulose content. Therefore our further studies will focus on the determination of spectral bands which are potentially sensitive for hemicellulose and lignin and the total lignocellulose content which are also important compounds of dry senescent plant materials.

References

Daughtry, CST, 2001, 'Discriminating Crop Residues from Soil by Short-Wave Infrared Reflectance' *Agronomy Journal*, vol. 93, pp. 125-131. doi: 10.2134/agronj2001.931125x

Daughtry, CST, Hunt Jr. ER, & McMurtrey III. JE, 2004, 'Assessing Crop Residue Cover Using Shortwave Infrared Reflectance' *Remote Sensing of Environment*, vol. 90 pp. 126-134. doi: 10.1016/j.rse.2003.10.023

Daughtry, CST, McMurtrey III JE, Nagler PL, Kim MS & Chappelle EW, 1996, 'Spectral reflectance of soils and crop residues' In: A.M.C. Davies & and P. Williams (Eds.): Near Infrared Spectroscopy: The future waves, pp 505-511. Chichester, UK: NIR Publications.

Elvidge, CD, 1988, 'Examination of the spectral features of vegetation in 1987 AVIRIS data' Proceedings of the First AVIRIS Performance Evaluation Workshop, Pasadena, *CA. JPL Publication*, 97-101.

Elvidge, CD, 1990, 'Visible and near infrared reflectance characteristics of dry plant materials' *International Journal of Remote Sensing*, vol. 11 pp.1775-1795.

Erdei, É, 2013, 'Investigation of the valuable parameters of silage maize and sweet sorghum genotypes produced by mutation and heterosis breeding, in terms of silage and bioethanol production. Doctoral thesis. University of Debrecen

Georging, HK. & Van Soest, PJ, 1975. Agricultural Hand Book. U.S.D.A., p. 379.

Granström, M, 2009, 'Cellulose derivates: Synthesis, Properties and applications' Academic Dissertation, University of Helsinki

Hepworth, DG, Bruce, DM, Vincent, JFV. & Jeronimidis G, 2000, 'The manufacture and mechanical testing of thermosetting natural fibre composites' Journal of Materials Science vol. 35 pp. 293 –298.

Klinke, HB, Lilholt, H, Toftegaard, H, Andersen, TL, Schmidt, AS & Thomsen, AB, 2001, 'Wood and plant fibre reinforced polypropylene composites' In: 1st world conference on biomass for energy and industry. James & James (Science Publishers), pp. 1082 –1085.

Lavanya, D, Kulkarni, PK, Dixit, M, Raavi, PR & Krishna, LNV, 2011, 'Sources of cellulose and their applications- A review' *International journal of drug formulation and research*, vol. 2 pp. 19-38.

Lee, YG, Jin,YS, Cha, YL, Seo, JH, 2017, 'Bioethanol production from cellulosic hydrolysates by engineered industrial Saccharomyces cerevisiae' Biosource Technology, vol. 228. pp. 355-361. doi: 10.1016/j.biortech.2016.12.042

Madsen, B, & Lilholt, H, 2003, 'Physical and mechanical properties of unidirectional plant fibre composites – an evaluation of the influence of porosity' *Composite Science Technology*, vol. 63 pp. 1265 –1272. doi: 10.1016/s0266-3538(03)00097-6

Nagler, PL, 1997, 'Plant litter spectral reflectance' Master's thesis. University of Maryland.

Nagler, PL, Inoue, Y, Glenn, EP, Russ, AL & Daughtry, CST, 2003, 'Cellulose absorption index (CAI) to quantify mixed soil-plant litter scenes' *Remote sensing of environment*, vol. 87 pp. 310-325. doi: 10.1016/j.rse.2003.06.001

Nagy, A, Riczu, P, Gálya, B & Tamás, J, 2014, 'Spectral estimation of soil water content in visible and near infra-red range' *Eurasian Journal of Soil Science*, vol. 3 pp. 163-171. doi: 10.18393/ejss.69645

R, Core, Team, 2016, 'R: A language and environment for statistical computing' R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Roberts, DA, Smith, MO, Adams, JB, Sabol, DE, Gillespie, AR & Willis, SC, 1990, 'Isolating woody plant material and senescent vegetation from green vegetation in AVIRIS data' Proceedings of the Second Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) Workshop, JPL Publication, vol. 90 pp. 42-57.

Roberts, DA, Smith, MO & Adams, JB, 1993, 'Green vegetation, non-photosynthetic vegetation, and soils in AVIRIS data' *Remote Sensing of Environment*, vol. 52. Pp. 255-269.doi: 10.1016/0034-4257(93)90020-x

Riczu, P, Bíró, G, Sulyok, E, Nagy, A, Tamás, J & Szabó, Z, 2012, 'Determination of chlorophyll content in case of peach leaf curl disease (*Taphrina deformans*) with spectral analysis' *International Journal of Horticultural Science* vol. 18 pp. 49-52.

Khalil, SRA, Abdelhafez, AA & Amer, EAM, 2015, 'Evaluation of bioethanol production from juice and bagasse of some sweet sorghum varieties' *Annals of Agricultural Science*, vol. 60 317–324. doi: 10.1016/j.aoas.2015.10.005

Van Soest, PJ, Robertson, JB & Lewis, BA, 1991. 'Methods for dietary fiber, neutral detergent fiber, and non-starch polysaccharides in relation to animal nutrition' *Journal of Dairy Science*, vol. 74 pp. 3583–3597. doi: 10.3168/jds.s0022-0302(91)78551-2

Wagner H & Kaltschmitt M. 2013, 'Biochemical and thermochemical conversion of wood to ethanol – simulation and analysis of difference processes' *Biomass Convers Biorefinery*, vol. 3 pp. 87–102. doi: 10.1007/s13399-012-0064-0

Watanabe TE, 2013, 'Introduction: potential of cellulosic ethanol' In: Faraco V, editor. Lignocellulose conversion: enzymatic and microbial tools for bioethanol production, Springer, Berlin doi: 10.1007/978-3-642-37861-4_1

Zhao, LY, Dolat, A, Steinberger, Y, Wang, X, Osman, A, & Xie, GH, 2009, 'Biomass yield and changes in chemical composition of sweet sorghum cultivars grown for biofuel' *Field Crops Research*, vol. 111 pp. 55–64. doi: 10.1016/j.fcr.2008.10.006