DECEMBER 2016 NO. 2

BAVARIAN JOURNAL OF APPLIED SCIENCES

www.jas.bayern


BAVARIAN
JOURNAL
OF APPLIED SCIENCES®

BAVARIAN JOURNAL OF APPLIED SCIENCES®

ISSN: 2366-3952
The Bavarian Journal of Applied Sciences is published annually.

CONTACT US

Bavarian Journal of Applied Sciences
Technische Hochschule Deggendorf
Edlmairstr. 6+8
94469 Deggendorf, Germany
Phone: ++49 (0) 991 3615-0
Fax: ++49 (0) 991 3615-297
E-Mail: info@jas.bayern
Web: www.jas.bayern

INHALT

VORWORT

Liebe Leserinnen und Leser,

für die zweite Ausgabe des Bavarian Journal of Applied Sciences (BJAS) gilt noch immer, dass es sich, ähnlich wie für die erste Ausgabe, um die ersten Gehversuche eines ehrgeizigen Projektes handelt. Im Unterschied zur ersten Ausgabe haben dieses Mal alle Beiträge ein Peer-Review-Verfahren durchlaufen. Zudem gibt es im aktuellen Heft zwei disziplinäre Schwerpunkte. Drei Beiträge sind den ingenieurwissenschaftlich-technischen Disziplinen zuzuordnen und drei Beiträge der Volkswirtschaft. Die ingenieurwissenschaftlich-technischen Beiträge befassen sich mit Methoden, die optische Prozesse und Messverfahren sowie Darstellungen durch Softwareanwendungen verbessern oder ermöglichen sollen. Die wirtschaftswissenschaftlichen Beiträge beschäftigen sich mit dem Zusammenhang von Preisen und Wechselkursen, den Erklärungsvariablen für Zinsentwicklungen sowie mit dem Widerspruch aus Intention und Wirkung von Gesetzgebung.

Der ingenieurwissenschaftlich-technische Abschnitt beginnt mit einem Beitrag von Gerald Fütterer aus dem Bereich Interferenztechnik. Fütterer diskutiert die Implementierung von Volumengittern, die auf Bragg-Beugung basieren und eine natürliche Wahrnehmung von 3D-Objekten in HD-Auflösung ermöglichen. Diese Methode ist für die Modellierung von 3D-Daten, für die Erstellung von Prototypen, CAD-CAM, Computer-integrierte Fertigung, weltweit arbeitende Telekonferenz-Systeme sowie für mobiles Infotainment von Bedeutung.

Ebenfalls der Disziplin Optik zuzurechnen, stellt der reproduzierte Beitrag von Engelbert Hofbauer, Rolf Rascher, Johannes Liebl, Manon Schilke und Jan-Peter Richters ein neu entwickeltes deflektometrisches Messverfahren vor, das es ermöglicht Krümmungen großer optischer Flächen mit einem einfachen Scanverfahren zu messen. Durch das vignettierende Feldblenden-Verfahren können auch Asphären und Freiformflächen ausgewertet und rekonstruiert werden.

Peter Faber, Stefan Schuster und Tanja Maier widmen sich der Wartbarkeit von Softwareprojekten und bewerten dafür Metriken als Methoden in der Softwareentwicklung. Anhand einer Anwendung untersuchen die Autoren den Beitrag der Metriken zur Verbesserung der Qualität eines InCarApp-Projektes innerhalb eines Forschungsprojektes zu Elektromobilität im ländlichen Raum. Die Reichweitenberechnung für Elektrofahrzeuge konnte erheblich verbessert werden und damit eine Eintrittshürde für Autofahrer zu Elektrofahrzeugen abgebaut werden.

Der wirtschaftswissenschaftliche Teil wird durch den Beitrag von Kersten Kellermann und Carsten-Henning Schlag eingeleitet, in dem die Auswirkungen sich verändernder Wechselkurse des Schweizer Franken auf Importpreise, Konsumentenpreise für importierte Güter und den Konsumgüterpreisindex in der Schweiz zwischen 1980 und 2014 analysiert werden. Die Berechnungen der Autoren zeigen, dass ein Großteil des Wechselkurseffektes durch den grenzüberschreitenden Handel absorbiert wird.

Harm Bandholz, Jörg Clostermann und Franz Seitz widmen sich den Zinsentwicklungen in den USA. Die Autoren zeigen anhand einer ökonometrischen Analyse, dass es trotz der zuletzt expansiven Geldpolitik der Federal Reserve (FED) eher klassische Faktoren wie Notenbankzinsen, Inflationserwartungen und die Auslandsnachfrage nach US-Staatsanleihen sind, die signifikante Erklärungskraft für die Entwicklung der Langfristzinsen besitzen. Auch das Anleihenkaufprogramm der

Dear Readers,

Being an ambitious project, the Bavarian Journal of Applied Sciences (BJAS) is still trying to find its feet in its second year of existence. Two refinements were made compared to the first issue. Firstly, the entirety of articles has undergone a peer review process before being published and secondly, the current issue can be divided into two sections by discipline: three articles belong to the fields of engineering and IT and three contributions are from economics. The contributions in engineering/ IT deal with methods aiming at improving and enabling optical processes and measurement techniques as well as visual displaying through software applications. The contributions in economics address the relationship between prices and exchange rates, explanatory variables for interest rate development as well as the discrepancy between legislative intent and effect.

The first section related to engineering and IT starts with two contributions dealing with new optical methods. Gerald Fütterer's article, pertaining to the area of photonics, discusses the implementation of Bragg diffraction based volume gratings which allows for a small form factor and high-definition, natural viewing experience of 3D objects. This method is relevant for modeling 3D data, prototyping, CAD-CAM, computer-integrated manufacturing (CIM) as well as for global 3D teleconferencing systems and mobile infotainment.

The reprint by Engelbert Hofbauer, Rolf Rascher, Johannes Liebl, Manon Schilke and Jan-Peter Richters presents a new optical method called vignetting field stop procedure. The method uses a deflectometric approach to acquire big optical surfaces which allows for measuring nearly any shape or form using a scanning routine. As even sign changes in the curvature may be detected, aspheres and freeform surfaces of any size can be evaluated and reconstructed with the aid of this technique.

The contribution by Peter Faber, Stefan Schuster and Tanja Maier is dedicated to the maintainability of software projects and to this end, they evaluate software metrics as tools in software development. The authors have accompanied an in-car app within the framework of a research project on electric mobility in rural areas and in this context, they examined the contribution of software metrics to the in-car app's quality. The app collects the car's situation during a trip and provides drivers with information on the infrastructure for charging stations as well as the range of the remaining battery level on a map. The software significantly improves the precision of state-of-charge/remaining range monitoring and contributes to the reduction of barriers to electric vehicle adoption.

This issue's second section on economics is introduced with a contribution by Kersten Kellerman and Carsten-Henning Schlag who study the exchange-rate pass-through (ERPT) in Switzerland between 1980 and 2014. They argue that the effects of changing exchange rates on import prices, on consumer prices for imported goods and on the consumer price index were not significant as the ERPT has been primarily absorbed by cross-bordering trade.

Harm Bandholz, Jörg Clostermann and Franz Seitz study the development of interest rates in the United States (U.S.) as a long-term phenomenon. Their econometric analyses show that despite the recent expansive monetary policy of the Federal Reserve System (FED), the development of long-term interest rates is still primarily

VORWORT

FED habe einen messbaren Einfluss auf die Langfristzinsen. Der zuletzt expansive geldpolitische Kurs der FED habe diese Zusammenhänge jedoch instabiler werden lassen.

Last but not least kommentiert Hanjo Allinger das 2014 verabschiedete Pflegeplanungsgesetz des Bundeslandes Nordrhein-Westfalen. Ziel des Gesetzes ist es, das Angebot für Pflegeinrichtungen besser zu steuern. Allinger kritisiert aus volkwirtschaftlicher Perspektive, dass eine Ausweitung des Angebotes für Pflegeplätze nicht notwendigerweise Sozialausgaben erhöhe, wenn angenommen werde, dass der Neubau dieser Einrichtungen auch steigenden Wettbewerb und damit Preissenkungen zur Folge haben könne. Fraglich sei zudem, ob die Intention der Gesetzgeber, vermeintlich unnötige Neubauten von Pflegeheimen zu unterbinden, überhaupt durch dieses Gesetz erreicht werden könne.

Den Reviewern sei an dieser Stelle für ihre Arbeit und ihre wertvollen Anmerkungen herzlich gedankt. Sie haben damit zur Einhaltung qualitativer Standards des Journals und zur Verbesserung der Qualität der Beiträge beigetragen. Hervorzuheben sind zudem die hervorragende editorische Arbeit Esther Kinateders und die technisch und gestalterisch gelungene Umsetzung der Ausgabe durch Sabrina Rockinger.

Das Herausgeberteam hat sich dazu entschlossen, die Reviewer auf der Website bekannt zu geben. Für das nächste Heft wird es eine weitere Profilschärfung geben, die darin besteht, dass ab dann alle Autorinnen und Autoren zunächst ein doppelt verblindetes Peer-Review-Verfahren durchlaufen müssen. Wird der Beitrag veröffentlicht, werden auch die Identitäten der Autorinnen und Autoren sowie der Reviewer offen gelegt. Zudem wird das Bavarian Journal of Applied Science künftig mit jährlich wechselnden thematischen Schwerpunkten erscheinen, die weiterhin aus verschiedenen disziplinären Blickwinkeln und mit unterschiedlichen Methoden bearbeitet werden können. Der Call for papers für die dritte Ausgabe 2017 erscheint zeitgleich mit der hier vorliegenden zweiten Ausgabe.

Für die zweite Ausgabe des Bavarian Journal of Applied Sciences wünschen Ihnen die Herausgeber aufschlussreiche Einblicke in die anwendungsorientierten Forschungsbereiche und eine anregende Lektüre.

influenced by traditional factors such as central bank interest rates, expectations on inflation rate and the foreign demand for U.S. bonds. Moreover, the FED's bond purchasing program does have a measurable impact on long-term interest rates. The authors state that these correlations, however, have become less stable in the light of the FED's recent expansionary monetary actions.

Last, but not least, Hanjo Allinger reviews the nursing law of the state of North Rhine-Westphalia, which was implemented in 2014 with the aim of better regulating the supply of nursing home beds. According to the author, first of all, stepping up the number of nursing home beds does not necessarily lead to an increase in social expenditures if it is assumed that newly-built nursing homes would also enhance competition, which could in turn entail price reductions. Secondly, he questions whether this law could at all achieve the legislators' intention to prevent the establishment of allegedly unnecessary new nursing homes.

The BJAS editorial team would like to specifically thank all reviewers for the time and energy spent on reviewing the contributions. Their invaluable constructive comments have helped to improve the quality of the articles and to keep up the journal's quality standards. Special thanks go to Esther Kinateder for her editorial work and to Sabrina Rockinger for the journal's technical realization.

In order to provide secure walking grounds, further changes will be introduced along with the third issue. First, forthcoming volumes will be published under an annually changing, specific topic that can be approached from different disciplinary and methodological angles. The call for the third issue, which will be published in December 2017, is announced alongside the current issue. Secondly, all articles will undergo a double-blind peer review. With the publication of this volume, reviewers' identities will be disclosed in a list on the journal's website.

The editorial team wishes you inspiring insights when reading this new issue of the Bavarian Journal of Applied Sciences.

BAVARIAN JOURNAL OF APPLIED SCIENCES®

Enabling Holographic 3D Displays with Bragg Diffraction Based Volume Gratings and First Approaches to the Reduction of Diffractive Cross Talk

Gerald Fütterer
Technische Hochschule Deggendorf

ABSTRACT

Photonics products often require enhanced optical functionality, which cannot be provided by state of the art optics. This is the case when high-end optical systems as for example holographic 3D displays are developed. Implementation of Bragg diffraction based volume gratings enables small form factor and high definition (HD) natural viewing experience of 3D objects. This is important for modeling 3D data, prototyping, CAD-CAM, computer-integrated manufacturing (CIM), global 3D teleconferencing and mobile infotainment.

For example, holographic 3D displays, which are based on space bandwidth limited wave segment reconstruction, profit from the specific characteristics of thick hologram gratings, which can be referred to as Bragg diffraction based volume gratings. It is explained how to adapt angular, spectral and polarization selectivity in order to provide specific optical functionality, which is mandatory in order to realize HD holographic 3D displays.

A short introduction to backlight units (BLU), complex valued spatial light modulators (C-SLM) and combined field lenses (cFL) of holographic 3D displays is given.
The description of Bragg diffraction based volume gratings is based on the coupled wave theory (CWT). The usage of the 2nd on-Bragg maximum modulation is described. The advantage of the 2nd Bragg diffraction order is explained. Polarization beam splitter (PBS) geometries are described. Specific layouts of holographic 3D displays are explained. BLU and cFL are evaluated. It is shown how the suppression of diffractive crosstalk can be obtained.

Oftmals benötigen neue Produkte der Photonik eine erweiterte optische Funktionalität, die nicht von optischen Standard-Komponenten bereitgestellt werden kann. Dies ist auch für die Entwicklung von holographischen 3D-Displays der Fall. Hierbei ermöglicht die Implementierung von Volumengittern, die auf der Bragg-Beugung basieren, eine kompakte Bauweise und eine natürliche Wahrnehmung von 3D-Objekten in HD-Auflösung. Dies ist für die Modellierung von 3D-Daten, die Erstellung von Prototypen, CAD-CAM, Computer integrierte Fertigung (CIM), weltweit arbeitende Telekonferenz-Systeme und mobiles Infotainment von Bedeutung.

Holographische 3D-Displays, die auf einer, in der Bandbreite limitierten Rekonstruktion von Wellenfrontsegmenten beruhen, profitieren von der spezifischen Charakteristik dicker holographischer Gitter, welche als Bragg-Beugung basierte Volumengitter bezeichnet werden können. Es wird beschrieben, wie die Winkel-, Wellenlängen- und Polarisationsselektivität ausgelegt werden kann, um spezifische optische Funktionalitäten bereitzustellen. Diese sind unter anderem notwendig, um HD-fähige holographische 3D-Displays zu realisieren.

Es wird eine kurze Einführung zu Hintergrundbeleuchtungseinheiten, komplexwertigen räumlichen Modulatoren und kombinierten Feldlinsen von holographischen 3D-Displays gegeben.

Die Beschreibung Bragg-Beugung basierter Volumengitter erfolgt unter Nutzung der CWT (coupled wave theory). Es wird dargelegt, wie das 2-te Maximum der on-Bragg Modulation genutzt werden kann. Der Vorzug der 2-ten Bragg-Beugungsordnung wird erklärt. Polarisations-Strahlteiler-Geometrien werden beschrieben. Spezifische Anordnungen von holographischen 3D-Displays werden erklärt. Beleuchtungseinheiten und kombinierte Feldlinsen werden untersucht. Es wird beschrieben, wie die Unterdrückung diffraktiven Übersprechens erreicht werden kann.

KEYWORDS

Bragg diffraction based volume gratings, holographic 3D displays, multiplex and diffractive cross talk

Bragg-Beugung basierte Volumengitter, holographische 3D-Displays, Multiplex und diffraktives Übersprechen

1. Introduction

The need for realistic 3D data visualization drives the increasing demand for displays which can provide all natural depth cues and thus natural 3D viewing experience. Unfortunately, there is no system on the market which can provide this. For instance, stereoscopic 3D, which is 2x 2D, can provide a limited comfort depth range ΔD_z only, which is approximately 15 % of the distance from the user to the display [1]. Furthermore, there are basic physical limits of these 3D display approaches used [2].

Holographic displays - still in the lab - and for example based on space bandwidth limited reconstruction of curved wave front segments, can provide all natural depth cues, a large depth range and a large viewing volume, too. Anyhow, they have not achieved commercial image quality and they are still not commercially available. Several questions may arise. Why does it take that much time? What are the bottlenecks and critical elements of the holographic display technology?

Real-time encoding of 1D-encoded display size spatial bandwidth limited holograms, which means, for example, the implementation of vertical parallax only (VPO), is not an issue any longer. This is done on high-end free programmable logic arrays (FPGA) already. Taking Moore's law into account, it can be assumed that real-time display size 2D encoding, including iterations for image quality optimization, of spatial bandwidth limited holograms, which means to use 2D-shaped sub-holograms, can be done on FPGA within the next decade. Thus, encoding is not the bottleneck.

Eye tracking of several users, which might, for instance, be based on triangulation, can be provided in real time already. Furthermore, modular arrangements of eye tracking units working in parallel and in a mutually

synchronized way can provide enhanced capabilities. Thus, eye tracking is not an issue.

C-SLM dynamically transform the wave field they are illuminated with. Wave fields are generated, which represent the 3D image [2]. Display size C-SLM providing real-time dynamic wave front transformation with a bit depth of at least 8 bit are not available. C-SLM are critical components. Liquid crystal (LC) based diffraction gratings (LCG), which make use of individually controlled electrode lines, can be used to redirect the wave field representing the 3D image to the eyes of the user. For example, the electrode pitch might be $\Lambda_{E\text{-}LCG}$ = 1.6 µm, which means to generate dynamic diffractive phase gratings with a minimum grating period of $\Lambda_{P\text{-}LCG}$ = 3.2 µm. Although display size attenuated phase shifting masks (APSM) had been introduced in high-resolution OLED panel production several years ago, LCG are still critical components. This holds for manufacturing, frame rate and image quality finally obtained. Consequently, it will take a few more years to develop C-SLM and LCG further in order to meet the final product specifications. Anyhow, C-SLM and LCG are not discussed herein in detail.

Holographic displays, which are based on space bandwidth limited reconstruction of curved wave front segments, require partial coherent illumination. This means to provide sufficient collimation and a reasonable large coherence length z_c. Both parameters are not an issue in standard display technology.

To use classic display size collimation of wave fields illuminating complex valued spatial light modulators (C-SLM), which provide wave field transformation, will result in bulky display embodiments. Array type collimation units, which might be used in order to reduce the thickness of the backlight unit (BLU), cannot provide high-definition viewing experience (HD), which is related to the angular - best case - resolution of the human eye, which is 1/60 ° deg.

Thick hologram gratings [3], which also can be referred to as Bragg diffraction based volume gratings, are key components, which enable small form factor holographic 3D displays [4, 5]. In this context, small form factor means that a display thickness of less than 0.1 times the display height can be realized. The main point is, that a wave transformation can be realized by using a thin foil only. A maximum efficiency of close to one can be obtained for this transformation, which corresponds to a single propagating diffraction order. Several geometries of reconstruction, which all might have a high diffraction efficiency, can be exposed into a single layer of holographic recording material, which might be, for instance, a photopolymer. Alternatively, several layers containing for example a single geometry of reconstruction only, can be laminated onto each other [6]. Thus, wavelengths and/or beam paths can be separated with high efficiency.

The specific functionality provided by Bragg diffraction based volume gratings is the reason why they are used within illumination units of small form factor holographic 3D display prototypes and as field lenses. Field lenses focus the light, which is modulated by the C-SLM, onto a plane in front of the eye of the user.

The task of the following analysis is to provide a tailored optical design and optimized parameter sets of the Bragg diffraction based volume gratings, which are used within holographic 3D displays, in a way that the diffractive cross talk is minimized.

2. Methodology and Analysis

In this section, a short overview of the methodology used is given. The path through the 3D display's components is described and analytic methods used are given. The analysis is carried out step by step, starting at the light source and finally reaching the users retina. Boundary conditions and basic assumptions are explained.

2.1 Generic holographic 3D display design, basic functionalities and basic boundary conditions

A generic refractive embodiment of a holographic 3D display is used to explain the function of the backlight unit (BLU), the C-SLM and the field lens (FL). The basic functions of the components can be found in folded diffractive versions of small form factor holographic 3D displays too.

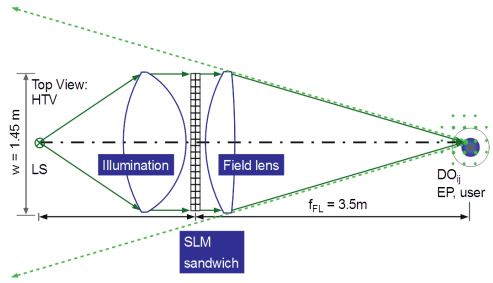


Figure 1: Generic refractive type embodiment of a holographic 3D display based on space bandwidth limited reconstruction of curved wave segments (Top view of a generic embodiment of a 65 inch display diagonal holographic TV, HTV, w: display width, f_{FL}: focal length of the field lens, EP: entrance pupil of the eye of the user, DO_{ij}: diffraction orders of the SLM in the EP plane). The dash-dotted black line is the optical axis. The dashed green lines indicate the addressable 3D viewing volume.

Figure 1 depicts a generic enfolded refractive embodiment of a holographic 3D display, which is based on space bandwidth limited reconstruction of curved wave segments. This principle will be explained within this chapter.

The light, which propagates from the light source, which is placed on the left-hand side, is collimated by using a best form lens, which is referred to as collimator or collimation unit. The sandwich type spatial light modulator (SLM) is illuminated with a reasonable plane wave.

In more detail, the angular spectrum of plane waves (ASPW), which has to be provided by the collimation unit, which is referred to as backlight unit (BLU) too, depends on the encoding used.

If 1D encoding and VPO is used, which is sufficient for holographic 3D TV displays, then the vertical ASPW propagating behind the SLM is $\Delta ASPW_v \leq 1/60$ ° deg and the horizontal one might be, for instance, within the range of 0.5 ° deg $\leq \Delta ASPW_h \leq 1$ ° deg. Note that the field lens function is superimposed to the VPO encoded lens functions. The circle of least confusion increases if the object point distance to the SLM plane is increased. The resulting astigmatism limits the HD range to approximately one third of the display-to-user-distance. For the generic embodiment being depicted in Figure 1 this limit is approximately 1.2 m in front of the display.

If 2D encoding is used, then the boundary condition to be fulfilled by the BLU is to provide an ASPW of ≤ 1/60 ° deg along the vertical direction and along the horizontal direction. The effect of an increased ASPW are smeared object points and thus a decreased depth range providing HD viewing experience, which is equivalent to the holographic reconstruction of 60 object points per 1 ° deg. It was assumed to limit the parts of 3D images, which are placed in front of the display, to 50 % of the user distance. A reduced ASPW has to be used in order to exceed this range.

It can be assumed that the sandwich type SLM is illuminated with an ASPW being tailored to the requirements of the discrete display embodiment.

The first SLM modulates the amplitude a(x,y,t). This is the A-SLM. The second SLM modulates the phase φ(x,y,t). This is the P-SLM. These two SLM are arranged very close to each other and form the C-SLM. A minimum propagation distance present between the A-SLM and the P-SLM is mandatory in order to reduce the diffractive crosstalk of adjacent C-SLM pixels.

For instance, the smallest clear pixel aperture defines the maximum distance d_{maxAP}, which might be present between the A-SLM and the P-SLM. Depending on the fill factor (FF) of the SLM pixels, the minimum clear aperture might for instance be 50 % of the SLM pixel pitch Λ_{SLM}. For a holographic 3D TV being placed at a user distance of z_{user} = 3.5 m, the upper limit of Λ_{SLM} is approximately 100 µm. Note that the SLM's diffraction pattern is present within the back focal plane of the field lens. This is depicted at the right hand side of Figure 1. The area present between adjacent diffraction orders is the area which can be used for the EP placement. For instance, using the grating equation, a wavelength of λ_B = 450 nm and a period of Λ_{SLM} = 100 µm results in a diffraction angle of 0.26 ° deg. At a viewing distance of z_{FL} = 3.5 m this results in a distance between the 0^{th} and a 1^{st} diffraction order of 15.75 mm. Taking the intensity distribution of the focal spot of the field lens and further deformations into account, which might be present, a viewing area of 10 mm x 10 mm is practical.
For Λ_{SLM} = 100 µm the clear pixel aperture might for example be 50 µm only. The analysis of internal diffractive SLM sandwich cross talk results in a maximum distance present between the A-SLM and the P-SLM, which is 10x of the minimum clear aperture. For this example, d_{maxAP} = 0.5 mm only. For this analysis, near field wave propagation was used. This was done by using the software MatLab and the wave

propagation module of the software Raytrace. This analysis will not be discussed herein. Consequently, one has to use very thin cover glasses with a thickness of less than 250 µm in combination with a face to face arrangement of the A-SLM and the P-SLM in order to get a C-SLM, which can be used for the holographic 3D TV consumer market.
The complex valued wave field, which is generated by the C-SLM sandwich type, can be described by:

$$c(x,y,t) = a(x,y,t) \times \exp(i(\varphi(x,y,t))) \qquad (1)$$

[2, 7]. This is realized by the C-SLM arrangement described above. Figure 2 depicts a generic visualization of a C-SLM formed by an A-SLM and a P-SLM in series. The first matrix represents amplitude values of the A-SLM and the second matrix represents phase values of the P-SLM. It can be assumed that this complex valued distribution has the size of a very small sub-hologram generating an object point being placed very close to the display plane. Note that the size of 2D sub-holograms, which generate real object points in front of the SLM plane, increases with increased distance present between the object points and the SLM plane. A real object point generated at 50 % of the display-to-user-distance requires an encoding area, which is at least as large as the EP of the user. At 25 %, the sub-hologram might be one third of the EP of the user only. In both cases, HD resolution can be obtained.

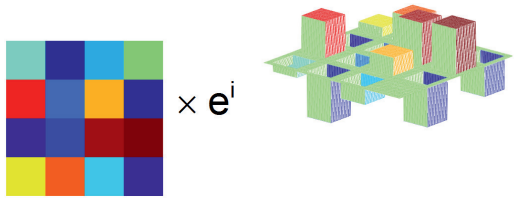


Figure 2: Generic description of the encoding of a complex valued wave field implemented by using a first amplitude modulating SLM (A-SLM) and a second phase modulating SLM (P-SLM) in series.

The complex valued wave field propagating behind the C-SLM represents the 3D scene. For example, single object points are represented by small local Fresnel zone plates generated by the C-SLM. Illuminating the C-SLM will generate object points in space. Depending on the sign of the radius of curvature R_{SH} of the curved wave segment provided by the local sub-hologram (SH), these are real object points in front of the display and imaginary object points behind the display.

Within holographic 3D displays based on space bandwidth limited reconstruction of curved wave segments, the light propagating behind the C-SLM has to be focused onto the user plane. This is done by the so-called field lens (FL). The field lens has a focal length f_{FL}, which is equivalent to the mean average user distance, which is for example f_{FL} = 3.5 m in the case of a holographic 3D TV (see Figure 1).

Within the focal plane of the field lens a diffraction grid type pattern is present, which is defined by the pixel pitch and the primary RGB wavelengths used. The area spanned between the 0th and a 1st vertical diffraction order (1D encoding, VPO) or the area spanned between the 0th, a 1st vertical and a 1st horizontal diffraction order (2D encoding) has to be placed in front of the entrance pupil of the users eye EP_{user}. This is the window to the pyramid type 3D viewing volume. The red, green and blue viewing windows of the three primary colors (RGB) have to have an overlay error of for example ≤ 1 mm. As it can be seen at the right hand side of Figure 1, the SLM can serve one EP only. In the embodiment discussed herein, both eyes are served time-sequentially. In addition, each color frame is generated time-sequentially. Each eye gets a different view. Thus, a 3D view presented to the user is made of six SLM frames.

Eye tracking has to be used in order to detect the position and the movement of the user eyes. A so-called tracking unit is required in order to redirect the wave field modulated by the C-SLM to the EP_{user}. Liquid crystal gratings (LCG) can be used for this [8]. For small movements two LCG, a vertical and a horizontal one, might be used. Within a display product at least three and up to five might be used for this task. Please note that the LCG are also used to correct aberrations, which will be present if diffractive wedge functions are added to a focus function of the field lens. The advantage of an LCG based tracking unit is that it is a thin embodiment already. However, a refractive backlight unit BLU enabling HD resolution is not thin. Furthermore, a refractive field lens (FL) enabling HD resolution is not thin either.

Using Fresnel lenses, surface relief type diffractive optical elements (DOE), as for

example Fresnel zone plates or multi order DOE, for the BLU or for the FL is explicitly not an option for HD consumer products. These are prototype solutions only, which generate unacceptable stray light. To be clear, these elements cannot be used as straight forward replacement of refractive or reflective BLU or FL. This is due to the basic physical properties of these optical elements and still holds even in the case of assuming perfect manufacturability. The entire chain of image forming, which also includes the ASPW used for the illumination of the C-SLM, has to be taken into account. Furthermore, multi-order DOE cannot provide the mandatory FL multiplex, which means to multiplex several FL geometries, for example two to seven, for each of the three wavelengths used.

Anyhow, the problem still is that the generic refractive type embodiment of a holographic 3D display based on space bandwidth limited reconstruction of curved wave segments, which is depicted in Figure 1, cannot be used for a mass product. Furthermore, Fresnel lenses or Fresnel zone plates cannot be used either. Thus, a further question may arise. Which type of optical elements can provide the mandatory optical functionality?
Bragg diffraction based volume gratings can provide the optical functionality which is required in order to realize small form factor holographic 3D display products, which are capable of generating HD object point clouds within a large 3D viewing volume. Thus, the problem of small form factor holographic 3D displays can be solved by implementing Bragg diffraction based volume gratings.

2.2. Practicability of Bragg diffraction based volume gratings

A basic boundary condition for the successful implementation of Bragg diffraction based volume gratings (BDVG) into holographic display products is to guarantee the manufacturability. This has to be done in regards to the specifications which have to be fulfilled in order to enable holographic 3D display products. Several aspects have to be taken into account.
Bragg diffraction based volume gratings are exposed into thin films of a recording material by using interference lithography. Based on classic holography, they are capable of transferring a reference wave into a signal wave and vice versa. Multiplex can be implemented within a single foil or within a stacked layer type arrangement. During the exposure or within a post exposure development process, the Bragg planes are generated. The diffraction efficiency η is chosen by generating the appropriate modulation of the refractive index n_1. A diffraction efficiency η close to 1 can be obtained. Note that the loss of optical energy, which is due to absorption and scattering, is excluded here. Parasitic diffraction orders can be totally suppressed. Several recording materials might be used, which can be offered in a reasonable price range. The materials can be used on glass or on plastic films as for example on triacetate cellulose (TAC) or other more expensive materials, like for example cyclic polyolefine, showing even better optical quality than TAC, which might be sufficient anyway.

Dichromated gelatin (DCG) requires a wet chemical development. High n_1 values of for example 0.1 can be realized with DCG. During the processing of DCG significant thickness variations Δd_{HOE} might occur. This results in a significant change of the geometry of reconstruction. For plane wave to plane wave geometries, a compensation might be implemented, for instance by changing the initial exposure angles. But the final angular reproducibility is within the 1 ° deg range, which is much too large for the application discussed herein.
Compared to DCG, photopolymers enable reduced angular deviations. The holographic recording film material HRF® (DuPont, Dai Nippon Printing) requires a post exposure bleaching and a post exposure bake at 120 ° C. Depending on the discrete material of the HRF® product family maximum values of n_1 of 0.05 or even up to 0.07 (fresh material only) might be realized. The shrinkage is still too large, although the shrinkage is less than the one present for DCG. Even with pre-compensation, the final angular reproducibility is within the 0.5 ° deg range [6], which is still slightly too large for the application discussed herein.
Newer photopolymers Bayfol® HX provided by the company Covestro are based on a matrix approach [9]. A primary polymer matrix provides form stability. The photopolymer system (monomers, oligomers and the photo sensitizer system) is placed within this matrix. Thus, very low post-exposure shrinkage is obtained [10]. The only post-exposure process

step is bleaching, which can be implemented by using high power white LED. Depending on the discrete material of the product family maximum values of n_1 of up to 0.04 might be realized. An angular reproducibility of < 0.1 ° deg can be obtained. Thus, the reconstruction is very close to the recording. This qualifies Bayfol® HX, for example placed on a TAC substrate carrier film, for the application discussed herein.

A further basic boundary condition for the successful implementation of Bragg diffraction based volume gratings into holographic 3D display products is to guarantee a reasonable low wave front distortion. From a global point of view, the light propagating behind the C-SLM has to be collected within a focal point ($\Delta x_F \leq \pm$ 0.5 mm). In other words, the FL acts as a global, display size lens. One has to ensure that the spot size of the focus of the FL, which is the zero order spot present in the plane of the EP of the user, does not exceed the 1 mm range. Low shrinkage type photopolymers can provide this. From a local point of view, the wave segment focused onto the retina has to generate a small point spread function (PSF). Aberrations excluded, the PSF is equivalent to the Airy disc [11]. The user should not detect an aberration of the object points. For single Bragg diffraction based volume grating foils, it is sufficient to work with aberrations of up to λ/8 at a 10 mm diameter. For photopolymers showing reasonable low shrinkage the reasonable low wave front aberration of the diffracted light was practically proven. In this context, the aberrations, which are due to thickness variations or stress birefringence of the carrier film, or due to local thickness variations of the photopolymer, have to be taken into account, too. Figure 3 illustrates that Bragg diffraction based volume grating can provide reasonable low wave front distortion, here < λ/60 at a diameter of Ø = 4 mm. For this measurement, a Shack-Hartmann sensor was used. The exit plane of the volume grating was imaged onto the detector plane.

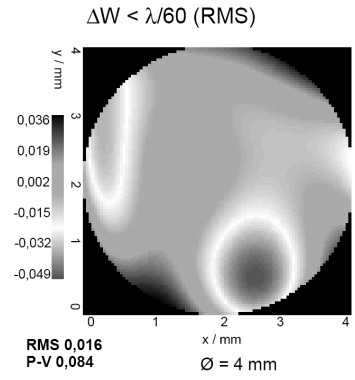


Figure 3: Measured wave front, diffracted by a volume grating (1st Bragg diffraction order [BDO], exit plane imaged onto the detector plane). The aberration is less than λ/60. The measurement area is equivalent to a representative area of the entrance pupil of the human eye EP (d_{EP} = 4 mm).

At this point, it can be concluded that Bragg diffraction based volume gratings can be used within holographic 3D displays, which are based on space bandwidth limited reconstruction of curved wave segment. In other words, this is a realistic candidate, which is not a solitary academic solution only. This means that Bragg diffraction based volume gratings are practically relevant. They are key components, which enable small form factor holographic 3D display products, which can provide HD-3D viewing experience.

2.3. Discrete implementation of Bragg diffraction based volume gratings

The bulky optical elements shown in Figure 1 have to be replaced by Bragg diffraction based volume gratings.

Figure 4 depicts an anamorphic diffractive type wedge BLU, which can be used in front of a C-SLM (in front: in respect to the beam path). The small form factor is obtained by using one dimensional anamorphic beam stretching two times, first along the horizontal direction and afterwards along the vertical direction. Thus, the first grating is a small stripe and the second grating has to have the size of the display.

If the BLU is used at 2x 87.13 ° deg / 0 ° deg (θ_{in} / θ_{out}), then 20x anamorphic stretching can be obtained. Anamorphic beam - or better wave field - stretching with a factor of 10x to 20x is practical. The anamorphic, diffractive wedge type backlight unit (BLU) can be realized, for instance, in polymethyl methacrylate (PMMA, used as wedge, for example with 20x beam stretching factor) or as air wedge (for example

with 10x - 15x beam stretching factor). The air wedge type BLU requires a dielectric multilayer coating optimized for the three primary wavelengths used (RGB, with a reflectivity of R < 5 % being practical at display size and 84.26 ° deg entrance angle). For instance, laser diodes with wavelengths of λ_B = 450 nm, λ_G = 520 nm and λ_R = 640 nm can be used. These wavelengths enable a wide color gamut. The final thickness of the BLU is approximately 1/10x to 1/20x of the display height. Thus, a small form factor BLU can be realized for the illumination of the C-SLM. Behind the BLU a narrow ASPW, which is close to a plane wave, can be assumed.

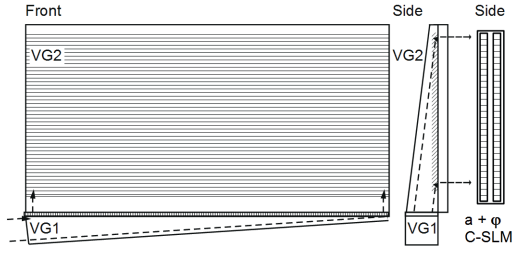


Figure 4: Left and middle: Front and side view of an anamorphic diffractive wedge type BLU. A collimated wave field propagates to a first stripe type volume grating (VG1), which is placed at the bottom. The entrance angle is θin = 84.26 ° deg and the exit angle is θout = 0 ° deg. This results in a 10x stretching factor present along the horizontal direction. The wave field, which is diffracted by the first VG, propagates to the second VG (VG2), which has the size of the holographic display. The entrance angle is θin = 84.26 ° deg. The exit angle, which is the angle the C-SLM is illuminated with, is 0 ° deg. Thus, two times 10x anamorphic stretching is obtained. Right: C-SLM being placed behind the BLU.

The C-SLM generates a spectrum of curved wave field segments, which are equivalent to a 3D cloud of object points. In order to generate the plane of the users viewing window to the 3D scene, which contains the diffraction orders of the C-SLM, a field lens is mandatory. In Figure 1, this is the lens being placed at the right-hand side of the C-SLM. This lens can be replaced by a Bragg diffraction based volume grating. Thus, a small form factor can be obtained. Furthermore, continuous phase profiles, which do not show any phase steps, are present. And last but not least, a multiplex of the geometries of reconstruction of three primary wavelengths and a multiplex of field lens geometries can be realized by using this approach. This enables an increased tracking range. One FL geometry, see for example Figure 5, has to be realized for three wavelengths. In other words, one FL geometry is R-G-B-multiplexed (R: red, G: green, B: blue). Coarse tracking can be combined with fine tracking in order to increase the freedom of movement of the user [8]. Several FL geometries can be multiplexed in order to implement coarse tracking. An important boundary condition, which has to be fulfilled for Bragg diffraction based volume gratings, is to realize a minimum diffraction angle, which is for example θmin ≥ 10 ° deg (Q factor [3]). Thus, an on-axis lens cannot be realized by using a single Bragg diffraction based volume grating. An on-axis lens can be realized by combining two Bragg diffraction based volume gratings in series. This is referred to as combined field lens (CFL). The on-axis CFL uses a plane-wave-to-plane-wave geometry of reconstruction followed by an off-axis illuminated on-axis focusing lens. This is depicted in Figure 5. Note that the maximum angle, which is realized by the field lens of a desktop type holographic display in air, is, for instance, up to 30 ° deg. Thus, it is sufficient to use an off-axis incident angle of for example 30 ° deg within the photopolymer (still: Q > 10).

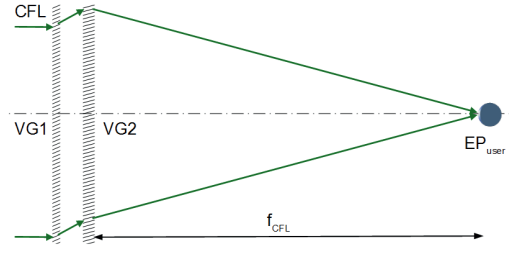


Figure 5: Principle of a combined on-axis field lens (CFL), which is based on two Bragg diffraction based volume gratings VG1 and VG2 in series.

Up to this point, the resulting beam path can be described as: collimated light source → BLU →C-SLM →CFL → fine tracking (several display size 1D LCG in series) → user being close to the focal plane FCFL of the CFL. In other words, the arrangement depicted in Figure 1 is made slim by using Bragg diffraction based volume gratings, which are tailored to the specific requirements of holographic 3D displays. For further analysis, the specific characteristics of the optical elements used has to be taken into account.

2.4. Angular and spectral selectivity vs. grating parameter

Several geometries of the reconstruction of Bragg diffraction based volume gratings had been taken into account. The specific geometries are results of specific optical designs of small form factor holographic 3D displays.

The calculation of the related diffraction efficiencies and the angular, spectral or polarization dependent change of them is based on the coupled wave theory (CWT) [3].
For the description of the Bragg diffraction based volume gratings, the selectivity of the diffraction efficiency η was analyzed in regards to a change of the entrance angle, the wavelength or the polarization state of light used for the reconstruction. The geometries of reconstruction, which had been used herein, are representative for holographic 3D displays. For instance, the 10x anamorphic air wedge uses a geometry of reconstruction of - 41.55 ° deg / 0 ° deg within the photopolymer, which means within a material that has a refractive index close to n = 1.5.

Figure 6 shows the calculated diffraction efficiency $\eta(\Delta\theta_B, n_1)$. Here, $\Delta\theta_B$ is the change of the entrance angle the Bragg diffraction volume grating is illuminated with in regards to the so-called on-Bragg angle θ_B. The value n_1 is the modulation of the refractive index of the phase only type of volume grating. The angular range shown is ± 5 ° deg. The modulation of the refractive index goes from zero to 0.04, which is the maximum value for time being Bayfol® HX photopolymers. The geometry of reconstruction within the photopolymer is - 41.55 ° deg / 0 ° deg. The thickness of the recording film is d_{HOE} = 20 µm.

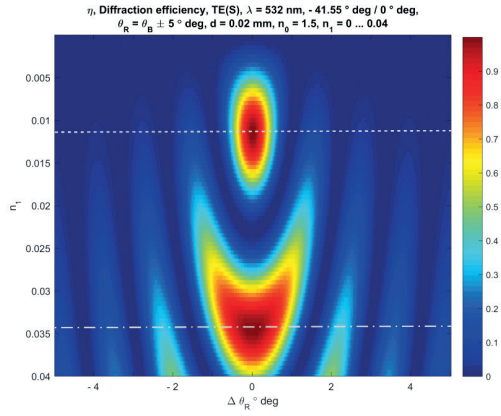


Figure 6: Calculated diffraction efficiency $\eta(\Delta\theta_B, n1)$ for an angular range of 10 ° deg and refractive index modulation n1 of up to 0.04.

The first on-Bragg maximum of the diffraction efficiency $\eta(\Delta\theta_B, n_1)$. is obtained at a modulation of n_{11} = 0.0115 (the 2nd index 1 indicates the 1st on-Bragg maximum of the function shown). This standard exposure state is indicated by a white dotted line. In other words, the standard situation is to realize an index modulation of n_{11} = 0.0115 for the related geometry and the related grating thickness. Slightly over modulating decreases the diffraction efficiency. Higher index modulation can be used to realize the 2nd on-Bragg maximum of the diffraction efficiency $\eta(\Delta\theta_B, n_1)$.

The second on-Bragg maximum of the diffraction efficiency $\eta(\Delta\theta_B, n_1)$ is obtained at a modulation of n_{12} = 0.0345 (the 2nd index 2 indicates the 2nd on-Bragg maximum of the function). This exposure state is indicated by a white dash-dot line.

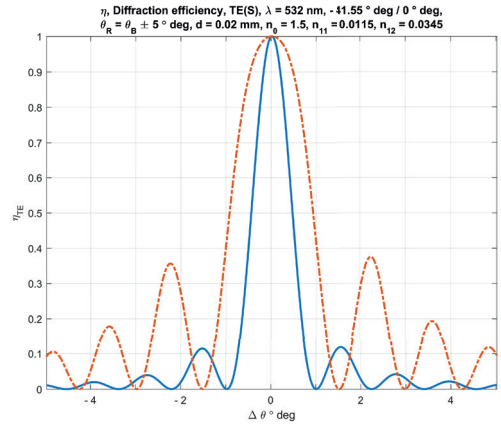


Figure 7: Calculated angular selectivity of the 1st (solid blue line) and the 2nd (red dash-dot line) maximum of the diffraction efficiency $\eta(\Delta\theta_B, n_1)$.

Figure 7 shows the calculated angular selectivity of the 1st and the 2nd maximum of the diffraction efficiency $\eta(\Delta\theta_B, n_1)$. The curves are cross sections of the function shown in Figure 6.

The point here is that the central angular range enabling diffraction efficiency $\eta(\Delta\theta_B) \geq 0.95$ is increased by a factor of 2.5 for the second on-Bragg maximum. This is of importance in the case of combining several Bragg diffraction based volume gratings together. On the one hand this means to be able to work with relaxed alignment tolerances, but on the other hand increased side lobes have to be taken into account. The analysis of the system described herein shows that this type of over modulation is not practical for the proposed system. As a result, the 1st on-Bragg maximum of the diffraction efficiency $\eta(\Delta\theta_B, n_1)$ has to be used. The geometries of reconstructions have to be optimized in regards to this fist maximum.

Figure 8 shows the calculated diffraction efficiency $\eta(\Delta\theta_B, n_1)$. Here, $\Delta\theta_B$ is the change of the entrance angle in regards to the on-Bragg angle θ_B. The value $\Delta\lambda$ is the change of the wavelength used for reconstruction. The geometry of reconstruction within the photopolymer is - 41.55 ° deg / 0 ° deg (BLU). The thickness of the recording film is d_{HOE} = 20 µm. The central wavelength is λ = 532 nm. The entire wavelength range used is ± 100 nm. Thus, the blue and the red spectral range can be found at the upper and at the lower rim respectively.

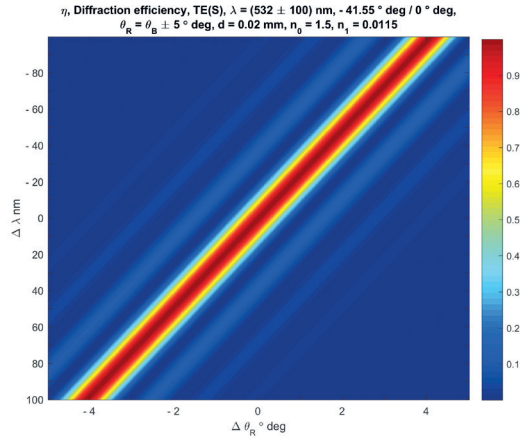


Figure 8: Calculated diffraction efficiency $\eta(\Delta\theta_B, \Delta\lambda)$ of a Bragg diffraction based volume grating designed for a λ = 532 nm BLU geometry.

In Figure 8, it can be seen that the green grating still shows very small ± 3rd side lobes, which are within the angular range of blue and red light. This holds if identical entrance angles are assumed, which might be realistic in the case of a wedge type BLU.

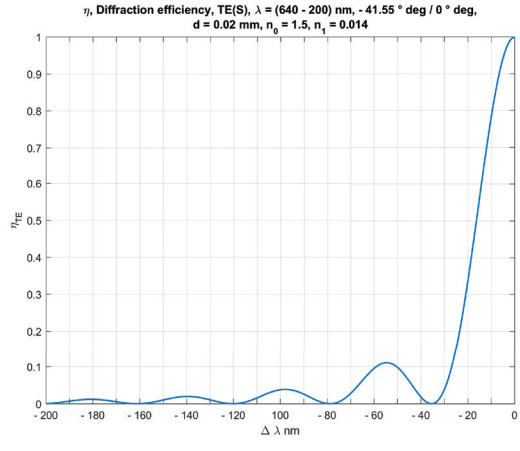


Figure 9: Calculated diffraction efficiency $\eta(\Delta\lambda)$ of a Bragg diffraction based volume grating designed for a λ = 640 nm BLU geometry at on-Bragg illumination.

Figure 9 shows the calculated diffraction efficiency $\eta(\Delta\lambda)$ for a λ = 640 nm BLU geometry at $\Delta\theta_B$ = 0 ° deg. The wavelength spans the range from λ_{max} = 640 nm to λ_{min} = 440 nm. Comparing Figure 9 with Figure 8 shows that diffractive crosstalk will be present in the case of multiplexing both geometries. The intensity of red light, which will of course be diffracted in a parasitic direction by the grating designed for 532 nm wavelength, is within the < 5 % range. The intensity of blue light, which will be diffracted in a parasitic direction by the grating designed for 640 nm wavelength, is within the < 2 % range. If the light does not propagate to the viewing zone of the user space, which is the case for the chosen geometry of diffraction, then it is just an acceptable loss of energy. The anamorphic BLU geometry is based on a plane-wave-to-plane-wave transformation. Thus, the intensity taken out of the beam path by the diffractive crosstalk is constant within the display plane.

The analysis shows that several options can be used in order to eliminate the diffractive cross talk within a wedge type BLU. The direct approach is to increase the grating thickness, for instance to a value of d_{HOE} = 30 µm. For this approach, the practical thickness limit is d_{HOE} = 40 µm. Figure 10 shows the calculated diffraction efficiency $\eta(\Delta\theta_B, \Delta\lambda)$. The geometry of reconstruction within the photopolymer is - 41.55 ° deg / 0 ° deg (BLU, air wedge, 10x). The thickness of the recording film is increased to the upper practical limit of $d_{HOE} \cong$ 40 µm. The central wavelength is λ = 520 nm, which enables a wide color gamut. The entire wavelength range used is ± 100 nm. The drawback of the increased thickness approach, which provides a much more narrow selectivity of the BLU gratings, is a much tighter alignment tolerance.

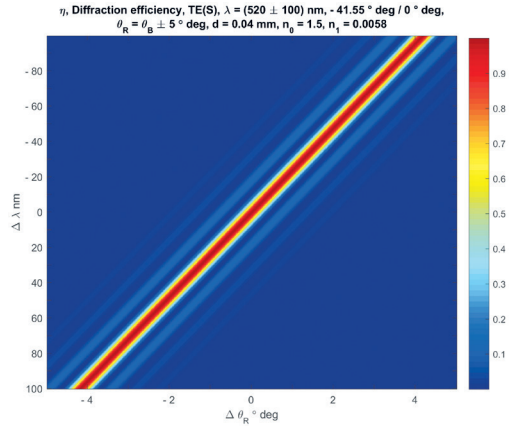


Figure 10: Calculated diffraction efficiency $\eta(\Delta\theta_B, \Delta\lambda)$of a Bragg diffraction based volume grating designed for a λ = 520 nm anamorphic BLU geometry (10x in air).

Another approach is to use an anamorphic BLU geometry of - 41.55 ° deg / 0 ° deg for λ_B = 450 nm and for λ_R = 640 nm, while using a BLU geometry of + 41.55 ° deg / 0 ° deg e.g. for λ_G = 520 nm. This means to flip the entrance angle of the green light.

To use λ_G = 520 nm instead of λ_G = 532 nm results in a wider color gamut. Reasons to use diode pumped solid state (DPSS) lasers working, for instance, at λ_G = 532 nm instead are the high optical power, the small line width, the small spectral drift and the high stability of the output intensity. However, a display product has to realize a wide color gamut. Thus, green lasers which meet all requirements have to be developed.

To use special, slightly changed wavelengths which are placed in the local minima of adjacent side lobes is not an option. The reversed way is more realistic: first define and fix the wavelengths, for instance to 450 nm, 520 nm and 640 nm and afterwards slightly change the geometries of reconstruction of the Bragg diffraction based volume gratings in order to get into the local minima of adjacent selectivity curves. For this, the angles and/or the layer thickness can be changed.

A further boundary condition is that 1D encoding uses a so-called incoherent direction. For VPO, this is the horizontal one. The ASPW along this direction is for example (0.5 to 1) ° deg. Furthermore, the coherent wave field modulated by the C-SLM has to be accepted by the Bragg diffraction based volume gratings within a reasonable wide angular range. This range is, for instance, ≥ ± 0.5 ° deg. In other words, the angular acceptance along the coherent direction has to be e.g. ≥ ± 0.5 ° deg. Gratings used behind the C-SLM of a 2D-encoded holographic 3D display have to accept ≥ ± 0.5 ° deg with reasonable high diffraction efficiency η, which is e.g. η ≥ 0.75, along the vertical and along the horizontal direction. For example, cutting out the small angular range of the user's viewing window to the 3D scene, which is the range between the 0th and a 1st diffraction order of the C-SLM grid, does explicitly not avoid higher diffraction orders in the focal plane of the CFL, which is the far field of the C-SLM, which contains the complex values representing 3D objects. To cut all higher orders off like this results in a huge loss of light. In the far field, the diffraction orders will be generated again. The practical approach to the suppression of higher diffraction orders is to use C-SLM pixel apodization like the well-known Gauss or Kaiser-Bessel windows and a reasonable wide angular acceptance angle of the Bragg diffraction based volume gratings.

2.5. Polarization selectivity vs. grating parameter

Another way which can be used in order to reduce diffractive cross talk is to take the polarization dependence of the diffraction efficiency η into account. Starting from the CWT [3], it can be shown [12] that the following equation holds for large angles of diffraction:

$$\nu_{TM} = \nu_{TE}\cos(\theta),\ \eta_{TM} = \sin^2(\nu_{TE}\cos(\theta));$$
$$\text{with } \nu = \pi d_{HOE} n_1/(\lambda\sqrt{(c_r c_s)}). \qquad (2)$$

The grating parameter ν contains the thickness dHOE of the Bragg diffraction based volume grating, the modulation of its refractive index n_1, the wavelength λ and the direction cosine of the reference and the signal beam c_r, c_s. Generalizing equation (2) and taking the separation of different polarization states into account leads to a set of polarizing beam splitter (PBS) geometries. This includes a 90 ° deg PBS geometry and a 60 ° deg PBS geometry, which can be used for beam combiner based C-SLM, which combine two adjacent phase pixels, which might be practical for ≥ 100 µm pixel pitch only. A diffraction efficiency of η_{TE} > 0.98 and η_{TE}/η_{TM} > 200 can be realized by using the first and the second Bragg diffraction order [6]. The second Bragg diffraction order has the advantage of using for example Λ_2= 600 nm grating period instead of Λ_1= 300 nm. This can be used to avoid decreased diffraction efficiencies which are due to a resolution limit of the material. Furthermore, relaxed exposure and measurement conditions can be obtained for total internal reflection (TIR) or close to TIR geometries.

Figure 11 shows the measured angular selectivity of a Bragg diffraction based volume grating PBS illuminated with TE and TM polarized light, which is optimized for the 2nd Bragg diffraction order.

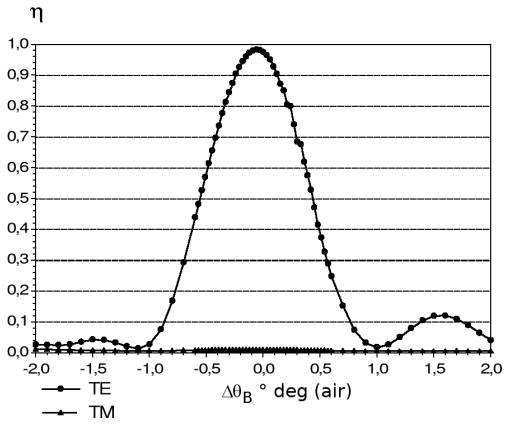


Figure 11: Measured angular selectivity of a Bragg diffraction based volume grating PBS illuminated with TE and TM polarized light. The diffraction efficiency is optimized for the second Bragg diffraction order (2nd BDO, η(Δθ) in air). The grating period is Λ_2= 600 nm, the diffraction angle is 90 ° deg within n = 1.5. The design wavelength and the wavelength used for reconstruction is λ_R = 650 nm.

A further analysis of equation (2) shows that there is a PBS geometry at - 41.41 ° deg / 0 ° deg, which is very close to the geometry of the 10x anamorphic air wedge BLU, which is - 41.55 ° deg / 0 ° deg. This implies to use a PBS geometry within the BLU, for instance for λ = 520 nm. A drawback of this approach might be the use of an additional multi-order retardation foil, which has to change the polarization state of the light if required. If C-SLM are used, which have color filters and patterned retarders for RGB, then this might be practicable. Using polarization selectivity for diffractive crosstalk reduction is an option. But it is not the preferred one.

2.6. Crosstalk within display size combined field lenses

The simulation of different geometries of reconstruction of display size components, which are characterized by a spatial variation of the local diffraction angle of Bragg diffraction based volume gratings, was carried out by using the optical design software Raytrace. This special optical design software was developed at the Chair for Optics of the University of Erlangen-Nuremberg.

Figure 12 is a central projection view of 3D ray tracing. A holographic desktop display with a focal length of f_{cFL} = 600 mm and an on-axis focus was assumed. The combined field lens used for this simulation is designed for 520 nm, but illuminated with all three display wavelengths. For this combined field lens, a first single grating was implemented, which diffracts all three wavelengths (see VG1 in Figure 5). This might for instance be a surface relief grating or a switchable LC grating. That is why the off-axis field lens designed for 520 nm wavelength is illuminated with RGB-angles, which are defined by the diffractive dispersion of the first thin grating. Thus, the parasitic colors generate deformed parasitic on-axis foci. Although the cross talk is low, the parasitic light will be focused within the user space. The analysis of the intensities shows that the foci are not that critical. There is other “on-axis” user at these positions. In this example, the green grating causes a large crosstalk and takes out light used for the blue and the red reconstruction. Thus, the simulation shows significant local intensity variations in the blue and in the red illumination planes. Intensity values within the range of 10 % of I_{mean} can be compensated by using a look up table (LUT). The local values of the LUT can be used as compensation values for the encoding of the local sub-holograms.

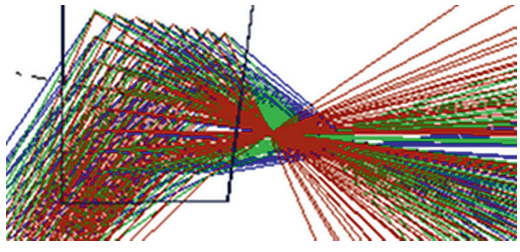


Figure 12: Raytracing of an off-axis Bragg diffraction based volume grating field lens, which is designed for λ_G = 520 nm. The lens is illuminated with λ_G = 520 nm, λ_B= 450 nm and λ_R = 640 nm wavelength. Collimated light enters the green off-axis field lens (comparable to VG2 in Figure 5) from the lower left-hand side. The lens is illuminated off-axis and forms an on-axis focus. The colors of the rays are equivalent to the wavelengths used.

An arrangement using propagation of RGB light at dispersion angles, which can be seen in Figure 12, is not preferable. This is due to the generation of significant parasitic diffractive crosstalk and intensity inhomogeneity. Thus, it is preferred to implement larger angles for shorter wavelengths and vice versa. This can be referred to as the implementation of reversed dispersion angles.

In order to provide a reasonable tracking range for the user, several field lenses have to be multiplexed for each color. It is preferred to use an LCG behind the C-SLM, which is only used for discrete angular switching. This is required to address a set of multiplexed field

lens geometries. This can be referred to as controllable coarse tracking. Behind the stack of multiplexed combined field lenses a stack of polarization type LCG can be placed, which provides the fine tracking.

An embodiment with reduced complexity might use two switchable off-axis field lenses for each color. The horizontal off-axis angle might for instance be ± 10 ° deg. This means to use six multiplexed combined field lenses in total. Twelve Bragg diffraction based volume gratings are used for this (mind the CFL). This points out that low cross talk, low scatter and low absorption are important parameters.

Figure 13 shows a top view of a Bragg diffraction based volume grating lens designed for 640 nm wavelength and a 10 ° deg off-axis focus. The focal plane is placed 600 mm in front of a holographic display. The lens is illuminated with λ_G = 520 nm, λ_B = 450 nm and λ_R = 640 nm wavelength. The three wavelengths propagate at reverse diffractive dispersion. Here 20 ° deg is used for red, 30 ° deg for green and 40 ° deg for blue light. The ray trace analysis of the Bragg diffraction based combined field lenses has shown that this minimizes the diffractive cross talk.

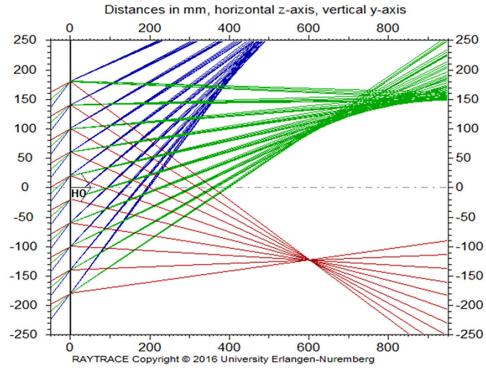


Figure 13: Top view of a Bragg diffraction based volume grating lens designed for λ_R = 640 nm and 10 ° deg off-axis focus. The focal distance is 600 mm. The lens is illuminated off-axis with λ_G = 520 nm, λ_B = 450 nm and λ_R = 640 nm. Illumination angles, which are reverse to diffractive dispersion, are used. The colors of the rays are equivalent to the wavelengths used.

Figure 14 shows the simulated distribution of the green (λ_G = 520 nm) intensity component of reconstruction at the exit plane of the red (λ_R = 640 nm) off-axis field lens. This field lens is the one shown in Figure 13. The maximum diffractive cross talk of the green light is 3.49 %. This is a reasonable low value. Note that there is no critical focus region formed and no zone of increased parasitic green intensity, which might disturb other users placed in the viewing range of the display.

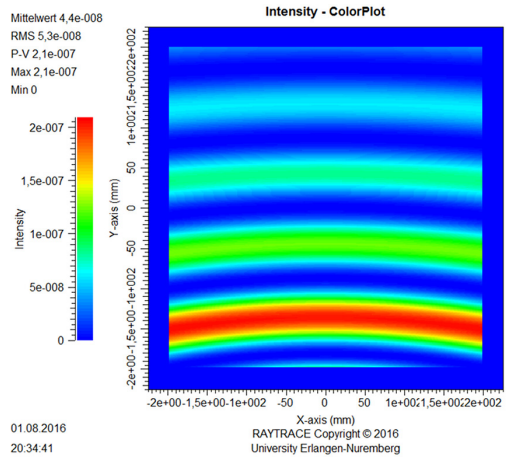


Figure 14: Distribution of the green (λ_G = 520 nm) component of reconstruction at the exit plane of the red (λ_R = 640 nm) off-axis field lens (see Figure 13). The lower part represents the left display side.

Due to the fact that the distribution shown in Figure 14 is stable, it can be directly implemented into a correction value type LUT. Thus, the results of simulations or even measurements (calibrations) can be used to correct the sub-holograms during the process of hologram encoding. In other words, one can take a realistic distribution of the illumination into account and thus increase the image quality of a holographic 3D display finally obtained.

Figure 15 shows the simulated distribution of the blue (λ_G = 450 nm) intensity component of reconstruction at the exit plane of the red (λ_R = 640 nm) off-axis field lens. This field lens is the one shown in Figure 13. The maximum diffractive cross talk of the blue light is 1.6 %. This is a reasonable low value. As it can be seen in Figure 13, there is no critical focus region formed which might disturb other users.

The question may arise how to interpret the distribution of the efficiency of the parasitic diffractive cross talk which can be seen in the Figures 14 and 15. This can be compared to Figure 8, which shows the calculated diffraction efficiency in the case of varying the design entrance angle and the design wavelength used for reconstruction. Note the side lobes. The simulations shown in Figures 14 and 15 are based on fixed entrance angles being not the design entrance angles and on fixed wavelengths, which are not the design wavelengths either.

Here, the local geometry of reconstruction is changed. Thus, side lobes of the diffraction efficiency can be seen. Furthermore, this is done in two dimensions, which is due to the field lens function. Thus, the parasitic cross diffraction type side lobes are curved. See also Figure 13 for orientation and note that this is a top view.

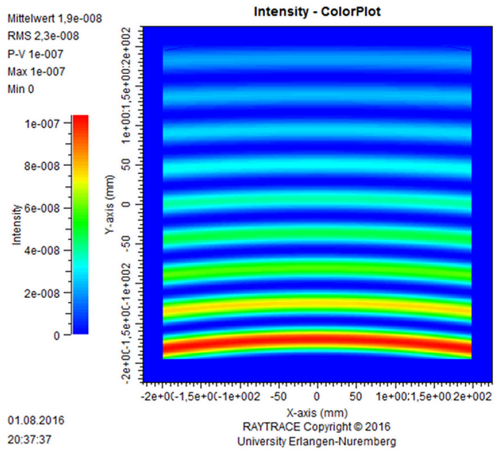


Figure 15: Distribution of the blue (λ_B = 450 nm) component of reconstruction at the exit plane of the red (λ_R = 640 nm) off-axis field lens (see Figure 13). The lower part represents the left display side.

The counter diffraction dispersive propagation of the wave fields of the different primary colors is a preferable approach in order to suppress the diffractive crosstalk within holographic 3D displays.

3. Results

The parameter space of Bragg diffraction based volume gratings provides reasonable high flexibility which can be used for the implementation of these gratings within holographic 3D displays.

The small form factor provided is practical for holographic 3D display product solutions.
The performance parameters of new photopolymers are sufficient for holographic 3D display product solutions, too.

The multiplex capabilities of Bragg diffraction based volume gratings can be used to realize the required beam paths for the primary display colors.

For coarse tracking, several geometries of reconstruction can be implemented for each primary color. Thus, it is practical to multiplex field lens functions.

The diffractive cross talk is a boundary condition, which has to be taken into account. The simulations had shown that film thickness and geometry variations can provide sufficient cross talk reduction.

In addition to this, the option of partial beam path flipping was evaluated. Thus, the mutual overlap of side lobes of the spectral and/ or angular selectivity curves of different diffractive geometries of reconstruction could be minimized.

The distribution of the diffractive cross talk is homogeneous for BLU geometries which are based on plane-wave-to-plane-wave transformations.

Combined field lens geometries could be optimized in a way to realize reasonable low diffractive cross talk. Residual intensity variations can be compensated by using a LUT. The values stored in a LUT are used in order to correct the encoding of the sub-holograms. Calibration data can increase the holographic 3D image quality finally obtained.
By using simulations, it has been shown that the diffractive crosstalk can be reduced to a practical level. Anyhow, further development has to be done within the next decades.

4. Conclusion

Bragg diffraction based volume gratings can be used within small form factor holographic 3D displays which are based on space bandwidth limited reconstruction of curved wave field segments. They can be used within anamorphic backlight units (BLU) providing, for example, up to 20x beam stretching. These types of diffractive BLU can provide collimated illumination of the C-SLM plane.

Bragg diffraction based volume gratings can be used in order to realize multiplexing of combined field lenses. The diffractive cross talk can be kept reasonable low. Although the manufacturing of large scale volume gratings which show small angular variations will be challenging, this initial cross talk study has pointed out that Bragg diffraction based volume gratings will be key components within small form factor holographic 3D displays which are based on space bandwidth limited reconstruction of curved wave field segments.

5. Outlook

Further optimization of the geometries of reconstruction, which are present within small form factor holographic 3D displays, can be carried out. This might also include slight shifts of for example the three primary RGB wavelengths λ_{01}, λ_{02} and λ_{03} used (but still at high color gamut). An angular flip, which might be used in order to minimize the diffractive cross talk, can, for instance, be carried out as up-down and as left-right operation. In other words, there are still additional options which can be used for cross talk reduction. Furthermore, longitudinal apodized volume gratings, which do not have diffractive side lobes at all, can be realized. This approach was theoretically confirmed by using Maxwell solver based simulations and it was practically confirmed by realizing tailored exposures [4]. This approach provides much denser multiplex operations than standard Bragg diffraction based volume gratings. This is important for providing a wide tracking range. However, much development has to be done in order to make display size longitudinal apodized volume gratings standard items.

6. Acknowledgement

The author would like to thank Prof. Dr. Norbert Lindlein of the Chair for Optics at the University of Erlangen-Nuremberg for providing the optical design software Raytrace, which is capable of analyzing the optical functionality of Bragg diffraction based volume gratings being integrated in classical optical systems.

7. Disclosure policy

The author declares that he has no financial interests with regards to this publication and that there are no conflicts of interests which might be related to this publication.

8. References

[1] Choi, Hee-Jin; Kim, Joo-Hwan; Park, Jae-Byung; Lee, Byoung-Ho (2007): "Analysis on the Optimized Depth of 3D Displays Without an Accommodation Error." In: The 7th International Meeting on Information Display (IMID 2007), Daegu, Korea; Paper 56-4, pp. 1811–1814.

[2] Poon, Ting-Chung (2006): *Digital Holography and Three-dimensional Display. Principles and Applications.* New York, London: Springer Science+Business Media Inc.

[3] Kogelnik, Herwig (1969): "Coupled Wave Theory for Thick Hologram Gratings." In: *Bell System Technical Journal* 48 (November), pp. 2909–2947.

[4] Fütterer, Gerald (2010): Compact Holographic Display Product Solutions. Invited Talk at the International Workshop on Holographic Memories and Display (IWHM&D). Tokyo, Japan, November 15–16, 2010.

[5] Fütterer, Gerald; Leister, N.; Haussler, R.; Reichelt, S.; Erler, C. (2011): Small Form Factor for Holographic 3D Display Product. Paper presented to 2011 Collaborative Conference on 3D & Materials Research (3DMR). Jeju City, South Korea, June 2011.

[6] Fütterer, Gerald: Holographisch optische Module in Photopolymer für DVD-Abtastsysteme. Diplomarbeit. Universität Erlangen-Nürnberg, Mai 1998. Lehrstuhl für Optik/Physikalisches Institut.

[7] Born, Max; Wolf, Emil (1991): *Principles of Optics.* Oxford: Pergamon Press.

[8] McManamon, Paul F.; Bos, Philipp J.; Escuti, Michael J.; Heikenfeld, Jason; Serati, Steve; Xie, Huikai; Watson, Edward A. (2009): "A Review of Phased Array Steering for Narrow-Band Electrooptical Systems. Rapid, Accurate, Non-mechanical Techniques for Steering Optical Beams Can Provide the Kind of Efficient Random Access Pointing Offered by Microwave Radars." In: Proceedings of the IEEE 1997,No. 6 (June), Vol. 97), pp. 1078–1096.

[9] Bruder, Friedrich-Karl; Deuber, François; Fäcke, Thomas; Hagen, Rainer; Hönel, Dennis; Jurbergs, David; Rölle, Thomas; Weiser, Marc-Stephan (2010): "Reaction-diffusion Model Applied to High Resolution Bayfol HX Photopolymer." In: Hans I. Bjelkhagen and Raymond K. Kostuk (Eds.): Proceedings of SPIE 7619. Practical Holography XXIV: Materials and Applications, pp. 761901I.

[10] Berneth, Horst; Bruder, Friedrich-Karl; Fäcke, Thomas; Jurbergs, David; Hagen, Rainer; Hönel, Dennis; Rölle, Thomas; Walze, Günther (2014): Bayfol HX Photopolymer for Full-color Transmission Volume Bragg Gratings. In: Hans I. Bjelkhagen and Michael V. Bove (Eds.): Proceedings of SPIE Vol. 9006. Practical Holography XXVIII: Materials and Applications, pp. 900602-10.

[11] Goodman, Joseph W. (1968): *Introduction to Fourier Optics.* New York: McGraw-Hill Book Company.

[12] Kobolla, Harald (1995): Holographische Komponenten für optoelektronische Datenvermittlungssysteme. Dissertation. Universität Erlangen-Nürnberg. Lehrstuhl für Optik/Physikalisches Institut.

Prof. Dr. Gerald Fütterer

Prof. Dr. Gerald Fütterer studierte Physik an der Technischen Universität Dresden und der Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU). Er promovierte über Interferenzlithographie bei λ= 157 nm und Interferometrie bei λ= 193 nm zur Messung von lithographischen Phase Shift Masken (PSM). Von 2004 an arbeitete er 3 Jahre lang an der Physikalisch-Technischen Bundesanstalt (PTB) in der Arbeitsgruppe Winkelmesstechnik der Abteilung 5: Fertigungsmesstechnik. Ein Schwerpunkt lag auf der Entwicklung optischer Messtechnik. Er führte PSM in Autokollimationsfernrohren (AKF) ein.
Nach Abwerbung durch die SeeReal Technologies GmbH in Dresden arbeitete er 9 Jahre lang an der Entwicklung holographischer 3D Displays. Er führte Bragg-Beugung basierte Volumengitter in holographische 3D Displays ein, baute interferenzlithographische Belichtungslabore auf (Master- und Kopier-Labor) und etablierte die Verarbeitung dieser speziellen Gitter. Ende 2008 wurde ihm die Position des „Head technology & Research“ angeboten, Ende 2011 dann die Position des Wissenschaftlichen Leiters (engl.: Chief Scientific Officer (CSO)). Er war maßgeblich an der Entwicklung holographischer 3D Displays beteiligt und als CSO für diese verantwortlich.
Seit 2016 ist er Professor für Produktionstechnik, Messtechnik, Berechnung optischer Systeme & Komponenten an der Technischen Hochschule Deggendorf.

Prof. Dr. Gerald Fütterer studied physics at Technical University Dresden and at Friedrich-Alexander-University Erlangen-Nuremberg (FAU). His dissertation was about interference lithography at λ = 157 nm and interferometry at λ = 193 nm being used for the measurement of lithographic phase shift masks (PSM). From 2004 onwards, and for three years, he worked at the Physikalisch-Technische Bundesanstalt (PTB, German NIST equivalent) in Brunswick within the working group of angular metrology, which is a part of the Division 5, Precision Engineering. He developed optical measurement systems and introduced PSM into autocollimators (AC).
After being poached by the SeeReal Technologies GmbH in Dresden, he developed holographic 3D displays for nine years. He successfully introduced Bragg diffraction based volume gratings into holographic displays and established processing of photopolymers by installing a master and a copy lab. In 2008, he was offered the position of Head technology & Research and in 2011 the position of Chief Scientific Officer (CSO). While serving as CSO, he led and was responsible for the development of holographic displays. Since 2016 he has been Professor for methods engineering, metrology, simulation optical systems & optical components at Deggendorf Institute of Technology (DIT).

✉ *gerald.fuetterer@th-deg.de*

DEC 2016 - NO. 2

BAVARIAN JOURNAL OF APPLIED SCIENCES®

Deflectometric Acquisition of Large Optical Surfaces "DaOS" Using a New Physical Measurement Principle: Vignetting Field Stop Procedure

Engelbert Hofbauer
Technische Hochschule Deggendorf
Technologie Campus Teisnach

Rolf Rascher
Technische Hochschule Deggendorf
Technologie Campus Teisnach

Manon Schilke
Technische Hochschule Deggendorf
Technologie Campus Teisnach

Johannes Liebl
Technische Hochschule Deggendorf
Technologie Campus Teisnach

Jan-Peter Richters
Berliner Glas KGaA
Dept. Messtechnik Entwicklung

ABSTRACT

The vignetting field stop procedure uses a deflectometric approach to acquire big Optical Surfaces – DaOS – and it offers the possibility to measure nearly any shape or form using a scanning routine.
The basic physical measurement principle in DaOS is the vignettation of a quasi-parallel light beam emitted by an expanded light source in auto collimation arrangement with a reflecting element. Thereby nearly any curvature of the specimen, is measurable. Due to the fact, that even sign changes in the curvature can be detected, also aspheres and freeform surfaces of any size can be evaluated. In this publication the vignetting field stop procedure is discussed. Additionally the deflectometric setup is described. Because of some typical influences of beam deflection to the accuracy of angle measurement by using the vignetting principle, suitable methods of calibration for the sensor are examined and the results of these methods are presented.

Furthermore, the technical principle of deflectometric measurements using an angle detecting device is explained inclusive of all random and systematic errors generated by the setup.

The last part of this publication shows the actual result of test measurements with calculated absolute deviation of errors with a large lateral dimension as well as the determination of the maximal achievable lateral resolution by detecting mid frequent structures on flat and spherical test parts with a diameter of 300 mm. These measurements are compared critically to reference results which are recorded by interferometry and further scanning methods.

Das Vignettierende Feldblenden-Verfahren bietet einen neuen, deflektometrischen Ansatz, um große optische Flächen mit nahezu jeglicher Form in einem einfachen Scanverfahren zu vermessen.
Das grundlegende physikalische Messprinzip in DaOS ist die Vignettierung der quasi-parallel von einer erweiterten Lichtquelle emittierten Strahlung, welche in Autokollimationsanordnung von einem reflektierenden Element zurückgestrahlt, abgeschattet auf einem Flächendetektor auftrifft und dort ausgewertet wird.
Damit ist nahezu jede Krümmung und auch ein Vorzeichenwechsel in der Krümmung des Prüflings messbar. Aufgrund dieser Tatsache lassen sich auch Asphären und Freiformflächen jeder Größe auswerten und rekonstruieren.

In dieser Veröffentlichung wird das Vignettierende Feldblenden-Verfahren kurz diskutiert und das deflektometrische Setup beschrieben. Einige typische Einflüsse der Strahlablenkung und geeignete Verfahren zur Kalibrierung des Sensors, welche die Genauigkeit der Winkelmessung beim Vignettierungsprinzip verbessern, werden untersucht und die Ergebnisse dieser Verfahren präsentiert.

Außerdem wird das Prinzip der deflektometrischen Anordnung mit Hilfe des winkelgebenden Sensors erläutert und die systematischen und zufälligen Fehlereinflüsse der Anordnung diskutiert.

Der letzte Teil dieser Publikation zeigt das tatsächliche Ergebnis von Testmessungen an Planflächen und Sphären bis 300 mm Durchmesser mit den berechneten absoluten Abweichungen der langwelligen Fehleranteile sowie die Bestimmung der maximal erreichbaren lateralen Auflösung an mittelfrequenten Strukturen.
Diese Ergebnisse werden mit Messungen interferometrischer Art sowie anderer Methoden verglichen.

KEYWORDS

Vignetting Field Stop Procedure, V-Spot, deflectometry, large optics, flats, spheres, aspheres, freeforms

Vignettierendes-Feldblenden-Verfahren, V-Spot, Deflektometrie, Großoptik, Planflächen, Sphären, Asphären, Freiformflächen

1. Introduction

In the field of precision optics mostly interferometry is used to determine shape errors of spherical or flat surfaces, but especially for large convex optics or aspheres, its functionality is strongly limited. For the deflectometric determination of any optical part using an autocollimator, the collimated beam is deflected through an invariant mirror system to the surface. The main advantage of the procedure is that no large and expensive references are

necessary so the systematical error, especially with large optics can be reduced dramatically.

Instead of a sharp picture, a mathematically clearly describable, defocused light spot is determined. The following image processing method calculates the value of the lateral offset corresponding to the angle of reflected beam.

2. New Measurement Principle

2.1 Deflectometric Flatness Reference DFR for flat surfaces

The principle of DFR is well known by using a scanning pentaprism [1] or penta mirror. Because of two internal reflections at the penta system, the deflected light bundle is invariant to angle deviations along the direction parallel to the Y-Axis –see figure 1. Roll angle errors influence the deflecting light in second order of roll angle. This works very well at flat surfaces up to the Sub-Nanometer scale [2,3].

The problem however is the measurement on strongly curved optical surfaces like spheres, aspheres or freeforms with slope changes larger than for example ± 1000 arcsec. The additional effect of vignetting by the classical autocollimator with crosshair or structure fixed in the object field as a point e.g. on the optical axis is not successful for such measurement, because reflected beams do not fit the aperture of the optic of the autocollimator. Therefore, this principle is limited by angle range and distance of light in optical path as shown in figure 1 in the right.

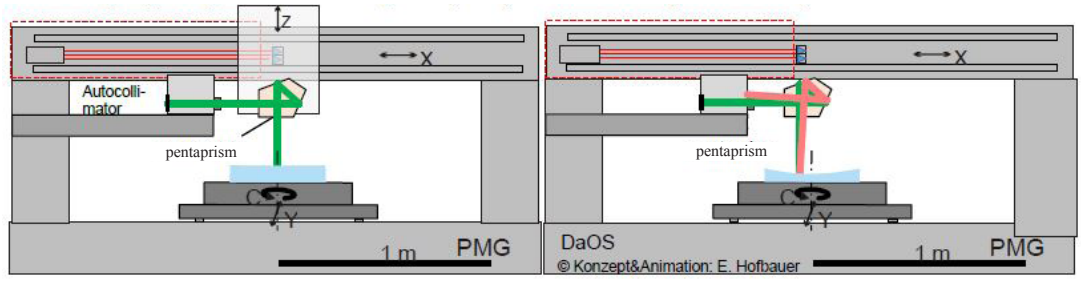


Figure 1. Setup of DFR on Precision Measurement Granite portal PMG. The emitted beam (green) from center of crosshair is reflected on Surface under Test SUT. For flat and smooth curves samples (radius in km-range), beam is reflected to the optics of autocollimator (left). In the case of higher curved surfaces with smaller radius (in m-range) reflected light cannot be detected because of failing the aperture of the optics (right).

2.2 Deflectometric acquisition of Optical Surfaces DaOS for large curved surfaces

Combining the **D**eflectometric **F**latness **R**eference with **V**ignetting **F**ield **S**top procedure [4] **VSP**, we come to the **D**eflectometric **a**cquisition of **O**ptical **S**urface **DaOS**. Therefor another data acquisition system is necessary. The Vignetting Field Stop or so called V-SPOT Procedure [4,5,6] is able to solve the problems of the classical autocollimator and gives further innovative solutions and options.

2.2.1 Vignetting Field Stop Procedure VFS

The Vignetting Field Stop VFS or so called V-Spot procedure is an special autocollimation principle using vignetting of the small aperture of the collimating optics. Instead of a fixed crosshair in object filed on the optical axis at the classical autocollimator, the larger emitting area allows varying the path of light, corresponding to perpendicularity of surface at specimen under test and illuminating of full entrance pupil. Figure 2 shows the setup of an autocollimation system with illumination and detecting area. The illumination area shows only 5 points in a meridional section of the setup. While failing the aperture of optics for the fixed light emitting point on the optical axis in case of tilted mirror (right side) light from the adjacent point (green rays) is reflected on the tilted mirror in retroreflection (rays are perpendicular to SUT) and is imaged on detector in the sensor system by nearly full intensity.

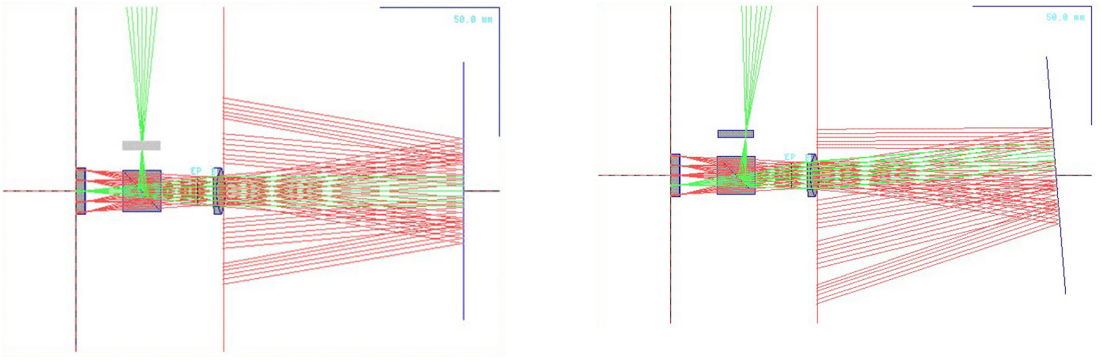


Figure 2: Setup of autocollimation system with illumination (only 5 from infinity points are shown) and detecting area. The rays that are perpendicular to reflector surface (green) are completely reflected back to system and give full illuminance in image plane.

Fig. 3. shows the image on a ½"-CCD-Sensor using a focal length of 140 mm with F# 5 at mirror distance of 500 (left) and 2000 mm (right).

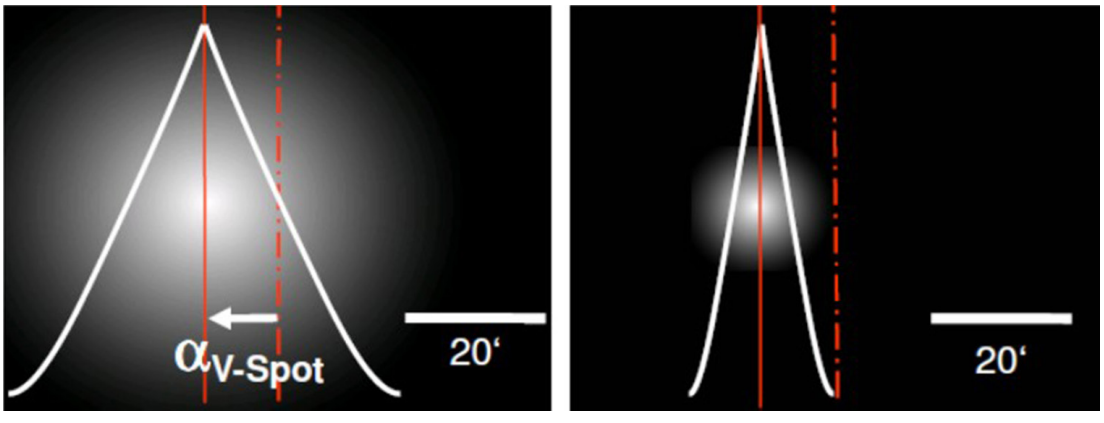


Figure 3: Image and Intensity distribution on an image sensor with 1/2" CCD-camera and focal length f' = 140/F#5

The image height y' is proportional double time to mirror angle α by the classical autocollimation setup known as

$$\tan 2\,\alpha = \frac{y'_{crosshair}}{f'} \qquad (1)$$

This means that the direct image of the crosshair is two time "faster" or two times larger than the pupil image which relates to one time of mirror tilt by

$$\tan\,\alpha = \frac{y'_{V-Spot}}{f'} \qquad (2)$$

This also means that the measurement range of equivalent sensors (same camera format and focal length) is two times larger at the Vignetting Field Stop VFS Sensor.

In practice, a focal length of 46 mm at F#5 with 1,2 MP-camera-format of about ½" is used. The physical resolution for movement detection in lateral displacement for the single binary V-SPOT detection is given by [7]

$$\Delta\alpha, \Delta\beta = \frac{1}{\pi \cdot \sqrt{r_{V-SPOT}}} \qquad (3)$$

with rV-SPOT is the radius of the binarised BLOB.

Even for a non-binarised V-SPOT the resolution will increase. Further research will be done and published in near future, to evaluate this limits. The accuracy over the range of ± 2,8° as shown in figure 4 is < 0,001° (3,6 arcsec) with

an average rms-value of < 0,00045° (1,6 arcsec) over a mirror distance range from 300 to 2.400 mm. As also seen in fig. 4 there is a limitation for linearity caused by the two dimensional calibration setup [6] whith PImiCos PRS-110 equipment.

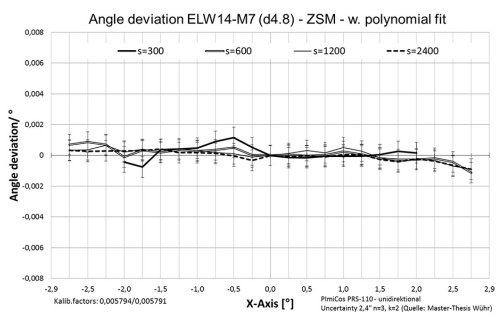


Figure 4: Result of the WiPoVi-Sensor calibration over a measurement range of ± 2.8 degrees at several distances from 300 to 2400mm. Error bars correspond to the uncertainty (k = 2, 1-α = 95%) of the angular reference table PImiCos PRS-110.

2.2.2 Deflectometric acquisition of Optical Surfaces DaOS with VFS

In order to detect also reflected light from the edge, VFS Sensor with angular measurement range up to slope on the edge, measurements can be done as seen in figure 5.

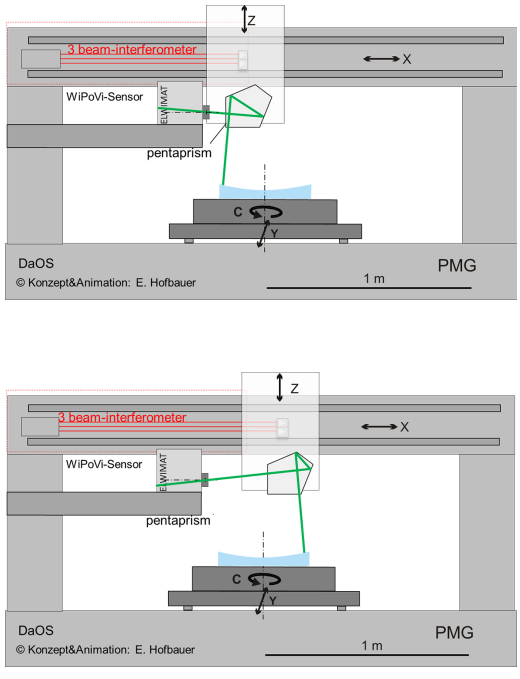


Figure 5. Setup of DaOS on Precision Measurement Granite portal PMG at Deggendorf Institute of Technology, shown in two edge positions. One of the emitted beam within the measurement range will always be perpendicular to the surface under Test SUT, and therefore will be reflected directly back to the sensor and will fit the aperture of the optics of the V-SPOT-Sensor.

2.2.3 Reconstruction of optical surface by line scan methods

In order to reconstruct the surface of a mirror substrate, several radial sections will be measured, using scans at azimuthal directions (e.g. 0°, 45° 90° and 135°, figure 6). The more scans, the better reconstruction will follow. The best strategy considering time, temperature variation and reconstruction accuracy will be find out in next time on our project DoSuRe. In addition to these radial sections, measurement in one or more circular sections are necessary to get correct reconstruction including all informations about aberrations of 3rd and higher order like Coma, Trefoil and others.

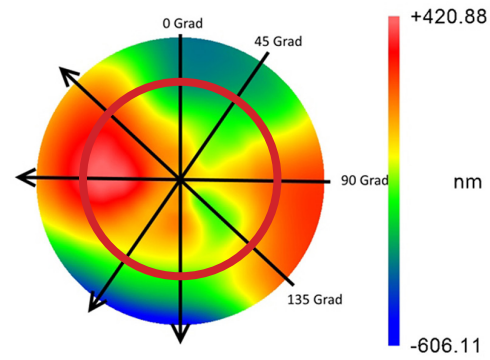


Figure 6. Lines and optional ringsegments of future DaOS-Surface evaluation of large optics

To reconstruct the complete surface out of the small amounts of data points of the measured traces along the surface several approaches were created and due to a program routine in our laboratory the measured traces along the surface were converted in mod-format. This file type is used for stitching the complete surface out of the measured datasets with the analysis unit of a TAYLOR HOBSON® measurement system. The results were saved as xyz-files to enable an alignment of the pv, rms- and power-values with the analysis software MetroPro®.

For our own approach the development environment MATLAB® is used. The main part of all developed analysis algorithms is the use of a Zernike Polynomial Expansion.

At least three algorithms are used to reconstruct the surface topography. One of them was created from Dr. Richters at the company Berliner Glas. Here, the aligned scans are linearly interpolated in polar coordinates and then rasterized linearly to xyz and native Zygo MetroPro binary-format

for easier comparison using MetroPro®. The others were developed on the Technologie Campus Teisnach. In figure 7 the current results of an evaluation with the analysis software MetroPro® are shown. The measured datasets are recorded by the ELWIMAT_AKF 46-4,8#17.

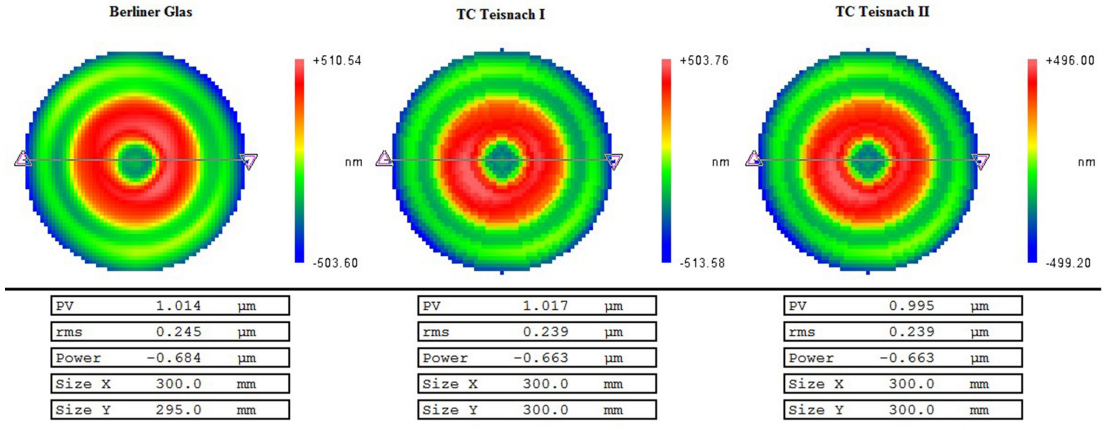


Figure 7: Alignment of the results of three developed evaluation algorithms in the analysis software MetroPro®.

The result shows maximum deviations in PV and power within 19 nm and 21 nm. The rms value differs at maximum 6 nm. With these first results, further investigations for the optimization of the algorithms are in progress in the following project DoSuRe. It is also seen that the simple and strong symmetric approximation and interpolation does not fit very well with interferometric and line scan measurements of DFR and DaOS. Further tests and theoretical simulations will follow.

3. Measurement Results

3.1 First results with single line scan

Very first results as shown in SPIE Proceedings [6,7] are done with the Sensor ELWIMAT 45-46 SN.#14, a Sensor with focal lenght of 46 mm by F# 4,8 and results of measuring linearity described in capture 2.2.1.

Sample one, a mirror with a smooth radius R = 20 m, Ø300 mm, coated with aluminum has been measured according DaOS with this sensor.

The result is shown in figure 8. The maximum slope at the edge belongs to 6,8 mrad (0.389°).

3.1.1 Radius evaluation with absolute accuracy

The sagitta height of 441,3 µm with an evaluated and calculated uncertainty of ± 0,5 µm at the measured diameter of 270 mm ± 0,7 mm enables to calculate the radius of curvature of 20.280 mm with uncertainty of 140 mm (k=2). This means a relative error of radius of 0,69% and a SAG error according to ISO 10110-5 of about 11 fringes (3/ 11 (-,-).

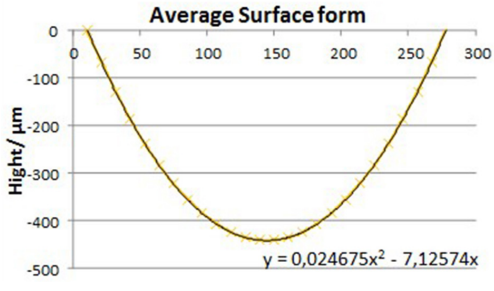


Figure 8. Average result of surface form deviation on mirror with radius 20 m.

3.1.2 Error budget at DaOS line scan

There are some influences in measurement errors which has to be considered and eliminated or reduced to get highest accuracy better than λ/10 @633nm. On the one hand there are systematic errors like geometrical deviations, penta prism adjustment and wavefront aberrations, linearity and also distance depending linearity effects of V-Spot sensor. On the other hand several random errors like temperature variations in measurement environment and on DUT, vibrations and air turbulences, random errors of sensor signal processing (signal-noise ratio) non reproducible roll angel and others may disturb the measurement.

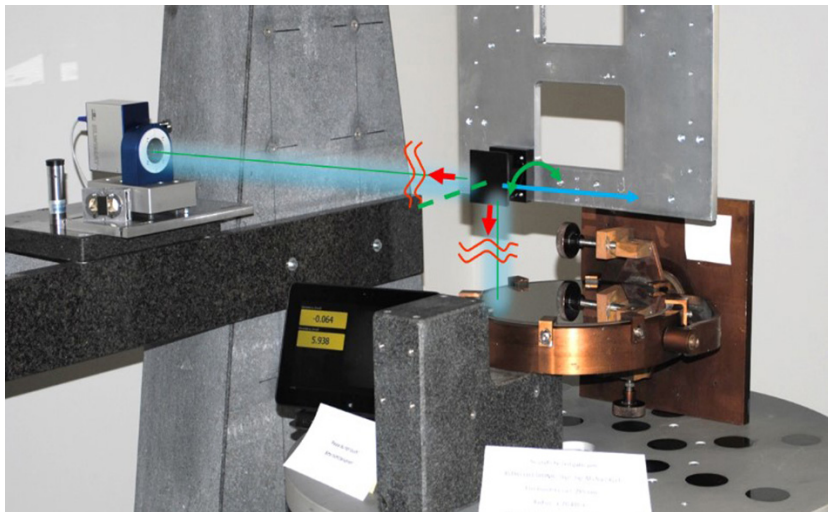


Figure 9: Measurement setup DaOS with radius 20 m mirror and some errors influences

Table 1: Error budget fpr the actual single line scan measurement setup at PMG with sample diameter 300 mm and step distance of 10 mm.

Influence in	Angle /arcsec	Hight/ nm
Random errors		
• Temperature difference on DUT up/down 70 nm/ 0,1K	0,2 (edge to edge)	70
• Vibrations and air turbulences	0,2	40
• Geometric/ adjustment errors penta prism < 2 arcmin	0,1	20
• Roll angle of X-Axis < 60 arcsec	0,02	4
• Random errors by V-SPOT Sensor noise	0,25	50
Systematic errors		
• Wavefront error penta prism 0,25λ/5mm	6	1.200
• V-SPOT Sensor long term non-linearity (focus, power) PV	3,6	200
• - short term deviations (coma, spherical, higher aberrations) rms - distance dependent V-SPOT Sensor error	1,6	45 50
Total statistical mean error		**1.250...1320 nm**

The greatest influences are seen in table 1 and are of systematic characteristic. First the wavefront error of penta prism with about 1200 nm (inhomogeneity of glass), second the long term sensor errors by non-linearity and third the distance dependent V-SPOT-Sensor error. To eliminate and evaluate this systematic influences, a reversal measurement is done measuring in direction forward and backwards with 180°-turned sample. The result seen in figure 10 shows a Peak to Valley at 410 nm (0,65 λ @633) and the rms about 112 nm (λ/6 @633).

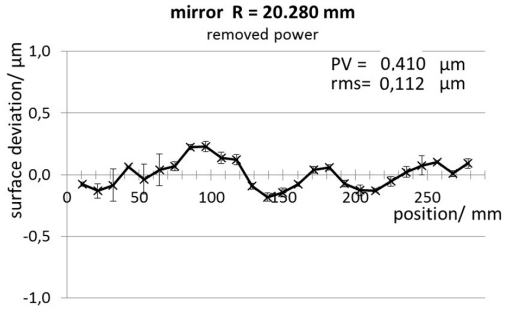


Figure 10: Surface deviation of mirror R 20.280 mm after correction of systematical errors by reversal measurement.

Compared with the first measurements described in SPIE-Proceeding 9132 913206 [6], without reversal measurement the deviation of about ± 1,5 micrometer (PV = 3 µm) is seen. This meets very exactly our consideration of above mentioned error budget.

3.2 Validation with a „Plano-Double-Sombrero"

In order to validate the measurement principle of new DaOS, a special "plano aspheric" geometry with "double sombrero" on a 300 mm diameter substrate was created (fig. 11) and measured by a short round robin test. R = plan; Ø = 308 mm, Thickness = 50 mm Øe = 300mm/ 294mm.

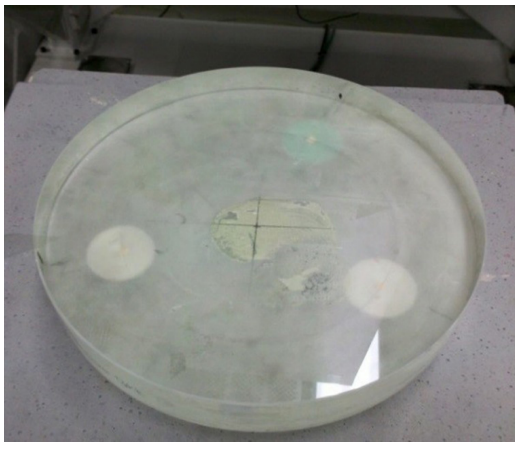


Figure 11: Plano-double sombrero with diameter 300 and thickness of 50 mm. Right: Crosscut in X- and Y-direction.

We used interferometry at the one hand and DFR in the other hand for evaluation. For Interferometry measurements a QED SSI-A instrument with a 4"-TF and a 6"-TF (both calibrated with tree flat test at TC Teisnach. Another evaluation is given by a 12"-Fizeau–Interferometer from ZYGO-Corp. in vertical position at cooperating company Berliner Glas. The DFR setup at PMG in Teisnach were done with three different classical electronic autocollimators:

- ELCOMAT 2000 (f=300)
- Trioptics Triangle (f=500)
- ELWIMAT_AKF 46/40 (f=46)

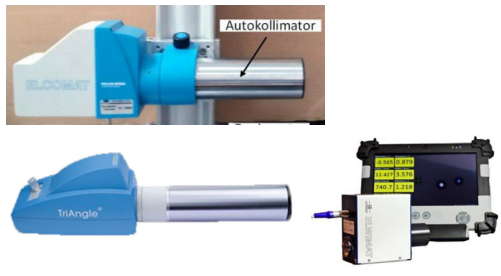


Figure 12: Instruments of electronic autocollimators for measurements DFR

At last measurements are done with the Vignetting V-SPOT Sensor ELWIMAT 45-46 #19 in order to evaluate this system in an absolute and reproducible way and also to evaluate measurement resolution in mid frequency structures to validate also coma and aberrations of higher order at polynomial function up to the 8th or 10th order.

3.2.1 Reference measurements with Interferometry

First we measured the plano-double-sombrero at SSI-A with 4"-TF transmission flat, which was calibrated using the tree flat test in 2012. Second and after some deviations on DFR measurements we made a three flat test on 4"-TF and also at a 6"-TF. Figure 13 shows the result of the 12"-Interferometer measurement at Berliner Glas and Table 2 includes the summarized PV, rms and SAG error (power) for all interferometric measurements.

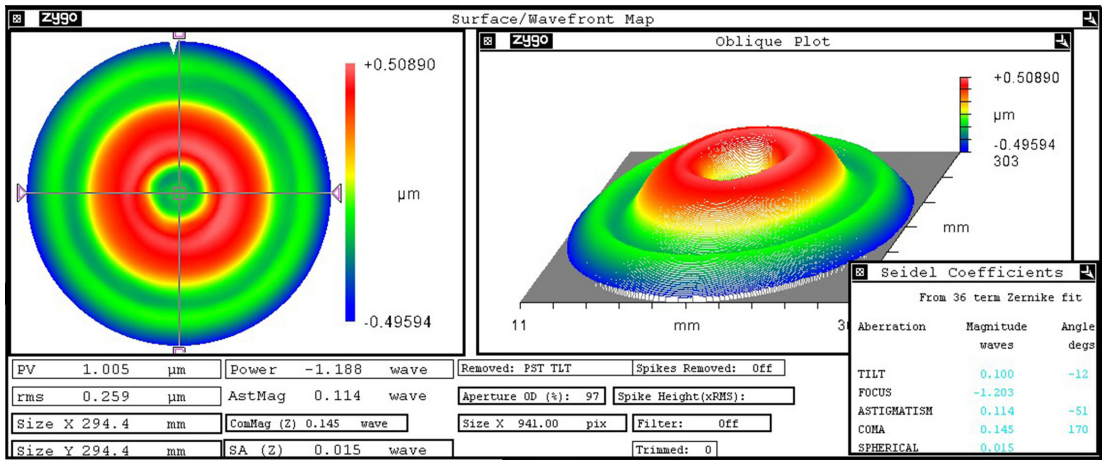


Figure 13: Interferometric measurement at plano-double sombrero, diameter 300, thickness 50 mm with 12" at Berliner Glas.

Table 2: Results in interferometric measurements of the „plano-double-sombrero" with Interferometric methods

	PV/ µm	Rms/ µm	SAG-Power/ µm
SSI-A 4" $Ø_e$ 294	0,955	0,249	-0,717
SSI-A 6" $Ø_e$ 294	0,973	0,244	-0,686
BG 12" $Ø_e$ 294	1,005	0,259	-0,752
Average	**0,985**	**0,251**	**-0,723**
Std. Dev.	0,037	0,008	0,039

The results shown in Table 2 offers an average PV-Value of 985 nm with a maximum deviation of 50 nm at all three measurements in two different laboratories and two different interferometer types. The absolute deviation of 3D-fitted power (SAG-error) is at maximum 66 nm. The standard deviation of the average of the three measurement results is 39 nm. The larger deviation of power at instrument at instrument at Berliner Glas is a gravity induced effect because due to the vertical orientation the transmission reference flat shows deformation of about 50 nm in the power term.

3.2.2 DFR-Measurement with electronic autocollimators

The deflectometric measurements with classical electronic autocollimators with crosshair were done in 4 sections at 0°, 45°, 90° and 135 degrees. A stop was applied near to the surface under test with an area of 14 x 20 mm for ELCOMAT 2000 by step 10 mm and diameter 5 mm for TriOptics 500mm and ELWIMAT_AKF 46 mm by step 5 mm. The measured scans are used to reconstruct the surface, feeding to software Taylor Hobson® for the ELCOMAT 2000 and the TCT-Algorithm with the other both autocollimators. The final results in PV, rms and power are seen in table 3.

Table 3: Results in DFR measurements of the „plan-double-sombrero" with electronic autocollimator

	PV/ µm	Rms/ µm	SAG-Power/ µm
ELCOMAT 2000 $Ø_e$ 294mm	0,965	0,253	-0,752
Trioptics AKF 500 $Ø_e$ 300mm	1,039	0,266	-0,793
EL WIMAT_AKF 46 $Ø_e$ 300mm	0,910	0,232	-0,640
Average	**0,974**	**0,250**	**-0,729**
Std. Dev.	0,061	0,017	0,079

Significantly we can see the PV-values which corresponds strongly with the influence of power in table 2 and 3, the quadratic term in the measured function. Varying the PV-value and power up to 129 nm and 153 nm, about three times more than at the interferometric measurements, we have to consider the influences at physical effects:

- Thermal drift of instrument holder during measurement
- Varying temperature difference of lower and upper side of specimen (table 1)
- Different sensitivity of instrument e.g. proportionality factor
- Linearity error of AKF and V-SPOT-instruments (table 1)

First we have to realise, that our environmental surroundings at PMG are not perfect because temperature compensation with about ± 0,5…1 K is not as good as in our laboratory (± 0,1 … 0,2 K).
Second, our "double sombrero" substrate is not quartz or ZERODUR® but probably Pyrex or similar material with a higher temperature expansion coefficient of about α = 3…4 E-6 1/K. This material is not the best choice, for high precision measurement of spherical influence (radius, power, SAG). In the case of pyrex we have calculated the influence of temperature gradient (upper to lower side of substrate) to vary by 70 nm per 0,1 K. Changing at the gradient up to 0,2 K, the Power or SAG error will rise up to 150 nm. In order to have no influence of algorithmic effects due to 3D reconstruction, we reduce data analysis to one dimensional 2D measurements for line scan as seen in figure 14. The spherical term is shown additional for two of the interferometric and the three autocollimation DFR measurements.

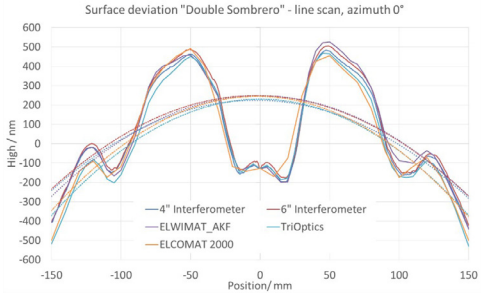


Figure 14: Interferometric and DFR measurements of plano-double-sombrero at azimuth 0°.

The graph in figure 14 shows the original measurements of interferometric and deflectometric DFR measurements in the azimuth direction of 0 degree. There is a good qualitative agreement at the characteristics of curvature of the structure of so called "double sombrero". But there is also seen, that the Deflectometric measurements varies strongly in the sagitta hight. The SAG errors (power) differs by about 150 nm, this is also seen in table 3. As also seen in figure 14, there are some decentering effects and the resolution varies between 10 mm (ELCOMAT 2000), 5 mm (ELWIMAT_AKF, Trioptics) and 0,89 mm for interferometric results.
In order to eliminate this temperature effects SAG error will be substracted from each measurement to get the Irregularity error. In figure 15 the average surface measurement of the two SSI-A interferometric measurements and the average of all three autocollimators is seen.
In order to compare the data and calculate the differencies of curves, the smaller amount of data in DFR and the following DaOS measurements by V-SPOT Sensor will be interpolated.

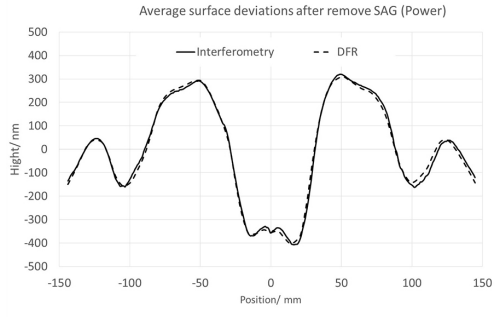


Figure 15: Averages of Interferometric and auto collimator DFR measurements at azimuth 0° after removing SAG error.

4. DaOS-Measurements on double sombrero with V-SPOT-Sensor

In order to evaluate the new V-SPOT-Sensor we use a ELWIMAT 46-4,8 UI1542#19 with focal length of 46 mm at F#4,8. Measurements are done in line scans at 0 degree of plane double sombrero.

4.1 Reverse measurement to reduce systematic influences

Most of the systematic errors like penta prism wavefront error and distance dependence error of the whole setup of DaOS are eliminated by reversal measurement.

The influence of these systematic errors as difference of the reverse measurements is shown in fig. 16. The value reaches a pv-value of 110 nm with an rms of 22 nm.

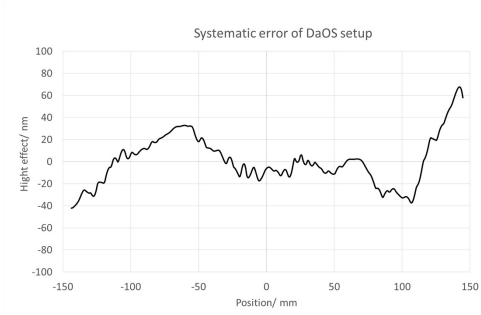


Figure 16: Evaluated systematic error of penta prism and V-SPOT sensor ELWIMAT 46-4,8 #19 by reverse measurement.

4.2 Validation of irregularity IRR after reducing SAG error (power)

To get elimination of above described effects we compensate the SAG error (power) at all measurements in order to reduce the influence of temperature effects during measurements and compare the results by viewing the so called irregularity IRR corresponding to the DIN-ISO 10110-5. The measurements were done by the modified and contrast enhanced V-SPOT Sensor ELWIMAT 46-4,8#19 in order to validate also non coated substrates with small reflectivity. With higher signal noise ratio we will be able to measure also at small aperture stops (scanning beam diameter) up to 1...2 mm or smaller.
The measuring result of DaOS with V-SPOT-Sensor is seen in figure 17. The table shows the very good agreement of all three measurement types including interferometry, DFR with auto collimator and DaOS with the Vignetting V-SPOT Sensor.

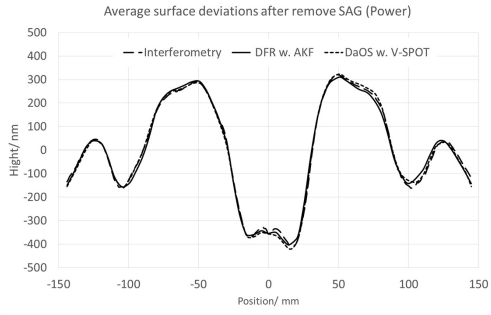


Figure 17: Interferometric, DFR- and DaOS- measurements of plan-double sombrero at azimuth 0°.

Table 4: Results in irregularity IRR (DIN ISO 10110-5) after removal of spherical part for all types of measurements at the „plane double sombrero“

	Interferometry	DFR w. AKF	DaOS w. V-SPOT-Sensor	Maximum dev.
PV = IRR/ nm	727	714	743	29
Rms/ nm	213,4	213,2	217,9	4,6

As described in Chapter 3.2.2, by interpolating of measurement data, we are able to take the residual curves of the measurements for an direct comparison. Therefore we take the difference of averaged DFR direct to interferometry and also DaOS to interferometry. The result shows, that there are residual errors of about pv 69 nm/ rms 14,5 nm at DFR with AKF and pv 63 nm/ rms 11,5 at DaOS with V-SPOT Sensor.

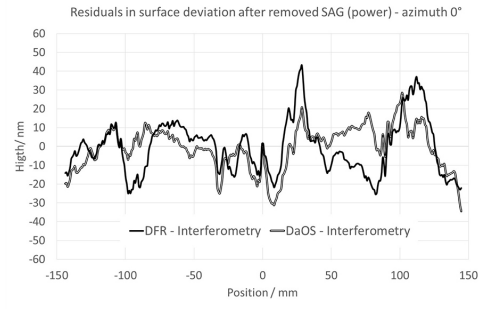


Figure 18: Residuals of DFR- and DaOS- measurements against interferometric results of plan-double sombrero at azimuth 0°.

There are some effects in the results in figure 18 visible, e.g. the large "peaks" at position 30 mm. A better adjustment of centration or an intelligent transformation of coordinates after measurement and bevor surface reconstruction may help to get better results than IRR < 60 nm according to λ/10 and rmsi < 11 nm according to λ/50 @633 nm even on large optics.

5. Summary/ Outlook

We could show, that it is possible to measure very large radius up to 20 m at large optics with more than 300 mm diameter. The relative accuracy of ΔR < 0,69 % will be reduced in further developments and future tasks.

Considering the sagitta error (SAG) we can see some temperature effects and influences of the quality of V-SPOT sensor calibration. Further on and considering the irregularity of spherical optics, we could show the measurements compared with interferometric stitching results at SSI-A and a 12"-Fizeau Interferometer at Berliner Glas, that it is possible to compensate systematical errors of higher order (nonlinearity and penta prism wavefront error) of the setup of DaOS. Good geometrical adjustments and also centration and compensation of mechanical deviations bevor surface reconstruction may help to get better results than IRR < 60 nm according to λ/10 and rmsi < 11 nm according to λ/50 @633 nm even on large optics.

This will be a good starting position even to have not a precision measurement room but measuring in a manufacturing environment.

6. References

[1] Debler, E.; Zander, K. (1997): Ebenheitsmessung an optischen Planflächen mit Autokollimationsfernrohr und Pentagonprisma. In: Physikalisch-Technische Bundesanstalt Braunschweig (PTB) (Hg.): PTB-Mitteilungen Forschen + Prüfen; 89. Jahrgang, S. 339–349.

[2] Weingaertner, I.; Elster, C. (2004): System of four distance sensors for high- accuracy measurement of topography. In: Precision Engineering 28 (2), S. 164–170.

[3] Lammert, H.; Noll, T.; Schlegel, T.; Siewert, F.; Zeschke, F. (2004): XUV-Optik – Neue unkonventionelle Bestimmung ihrer Flächenformen bei BESSY mit der NOM. In: 105. Jahrestagung der Deutschen Gesellschaft für angewandte Optik (DGaO) (01.-05.06.2004 in Bad Kreuznach).

[4] Hofbauer, E.; Lenz, M. (2004): Vignettierung als physikalisches Messprinzip. In: tm - Technisches Messen 75 (3), S. 153–167.

[5] Hofbauer, E. (2007): Optisches Verfahren zur zweidimensionalen Messung von Winkeln und Winkeländerungen bei großen Messabständen. Dissertation zur Erlangung des akademischen Grades Doktoringenieur (Dr.-Ing.). Technische Universität Ilmenau. Fakultät für Maschinenbau.

[6] Hofbauer, Engelbert; Rascher, Rolf; Stubenrauch, Thomas; Wühr, K.; Friedke, Felix; Pastötter, B. et al. (2014): 3D-optical measurement system using vignetting aperture procedure. In: Proceedings of SPIE 9132. Optical Micro- and Nanometrology V, S. 913206–913214.

[7] Zellner, T. (2005): Programmierung, Simulation und Optimierung des videogestützten optischen Winkelmesssystems ELWIMAT. Diplomarbeit. Hochschule München.

Dr. Engelbert Hofbauer

Engelbert Hofbauer studied Precision Engineering at Hochschule München and graduated as Dipl.-Ing. (FH) in 1985. He received his doctoral degree in Optical Engineering at Technical University of Ilmenau in 2008. From 2009 until 2013 he has held a research professorship at Deggendorf Institut of Technology. In this function, he built up the Laboratory for Metrology of Precision Optics Manufacturing at Technology Campus Teisnach.
Since 2013, he has worked as a researcher and project manager in applied research projects with the industry (WiPoVi, DoSuRe) on interferometric and deflectometric data acquisition in order to measure and reconstruct very large optical surfaces of at least plane, spherical, aspherical and also freeforms. In parallel, he directs the company Hofbauer Optik in Munich, which develops and manufactures autocollimation sensors and electronic sighting telescopes for precision metrology in engineering and for the optical industry (e.g. centration for spheres and aspheres).

Dr. Engelbert Hofbauer studierte Feinwerktechnik an der Fachhochschule München und schloss 1985 mit Dipl.-Ing (FH) ab. Er promovierte dann 2008 an der Fakultät Maschinenbau im Fachgebiet Technische Optik/Gerätetechnik an der Technischen Universität Ilmenau. Ab 2009 baute er während einer Professur den Messraum für Fertigungsmesstechnik am Technologie Campus Teisnach der Fachhochschule Deggendorf auf und leitete ihn bis 2013. Seither ist er als wissenschaftlicher Mitarbeiter und Projektleiter in anwendungsnahen Forschungsprojekten mit der Industrie (WiPoVi, DoSuRe) tätig und beschäftigt sich dabei mit interferometrischen und deflektometrischen Messmethoden zur Rekonstruktion großflächiger Optiken, von Planflächen bis Freiformflächen. Parallel dazu leitet er die Fa. Hofbauer Optik Mess-&Prüftechnik in München, welche Autokollimations- und Fluchtmesssensoren für die Präzisionsmesstechnik im Maschinenbau und für die optische Industrie (u.a. Zentriermessung für Sphären und Asphären) entwickelt und fertigt.

Kontakt / *Contact*

engelbert.hofbauer@th-deg.de

Prof. Dr.-Ing. Rolf Rascher

Prof. Dr.-Ing. Rolf Rascher ist seit 2001 Professor an der Technischen Hochschule Deggendorf, seit 2012 an der Fakultät Naturwissenschaften und Wirtschaftsingenieurwesen. Schwerpunkte der Lehre sind Fertigungstechnik und Technologiemanagement.
Seit Oktober 2009 leitet er den Technologie Campus Teisnach und die Forschungsgruppe „Optical Engineering" an der Technischen Hochschule Deggendorf. Daneben betreut er eine Vielzahl von Industrieprojekten aus dem Bereich der Fertigungs- und Messtechnik Optik. Forschungsschwerpunkte sind die Verfahrens- und Prozessentwicklung im Bereich der Optikfertigung und der Bearbeitung sprödharter Materialien.
Prof. Dr.-Ing. Rascher ist Hauptvorsitzender bei der jährlichen SPIE Konferenz „European Seminar on Precision Optics Manufacturing" am Technologie Campus Teisnach.

Prof. Dr. Rascher has held a professorship at Deggendorf Institute of Technology since 2001. Since 2012, he has been a professor at the faculty of Natural Sciences and Industrial Engineering. His focus of teaching is in the areas of production technology and technology management.
Since October 2009, he has been head of Technology Campus Teisnach and DIT's research group „Optical Engineering". He takes care of a variety of industrial R&D projects in the area of manufacturing and measuring technology for optics. The main focus of his research work is the process development in optics manufacturing and the treatment of hard, brittle materials.
He is first chairman of the annual SPIE conference „European Seminar on Precision Optics Manufacturing" held at Technology Campus Teisnach.

Kontakt / *Contact*

✉ **rolf.rascher@th-deg.de**

Manon Schilke, B. Eng.

Manon Schilke hat an der Technischen Hochschule Deggendorf (THD) ein Bachelorstudium in Physikalischer Technik mit dem Studienschwerpunkt Sensorische Systeme absolviert. Aktuell ist sie am Technologie Campus Teisnach als technische Mitarbeiterin beschäftigt. Ihr derzeitiges Aufgabengebiet im Bereich der optischen Messtechnik umfasst unter anderem Softwareentwicklung und Datenanalyse.

Manon Schilke has earned a Bachelor's degree in Applied Physics with a focus on sensor systems at Deggendorf Institute of Technology (DIT). At the moment, she is a technical staff member at Technology Campus Teisnach. Her current area of responsibility in the optical metrology field includes amongst others the development of software and the respective data analyses.

Kontakt / *Contact*

manon.schilke@th-deg.de

Dipl. -Ing. (FH) Johannes Liebl

Johannes Liebl studied Electrical and Information Technology at the Fachhochschule Deggendorf until 2010. As early as during his internships and his diploma thesis he gained considerable experience in the field of optics. Right after completing his studies he started working as a project engineer at Technology Campus Teisnach. The focus of his work lies in optical manufacturing and metrology.

Johannes Liebl hat an der Fachhochschule Deggendorf bis 2010 Elektro- und Informationstechnik studiert. Bereits bei Praktika und in seiner Diplomarbeit konnte er zahlreiche Erfahrungen im Bereich der Optik sammeln. Im direkten Anschluss an sein Studium war er als Projektingenieur am Technologie Campus Teisnach tätig. Die Schwerpunkte seiner Arbeit liegen in der optischen Fertigungs- und Messtechnik.

Kontakt / *Contact*

johannes.liebl@th-deg.de

Dr. Jan-Peter Richters

Jan-Peter Richters received his doctoral degree in physics from the University of Bremen in 2010. Since 2012, he has been working in the metrology development department of Berliner Glas on interferometric data acquisition and computer-based data evaluation. His current scope of work lies in the refinement of interferometric measurement capabilities at Berliner Glas.

Dr. Jan-Peter Richters promovierte 2010 im Fachbereich Physik an der Universität Bremen. Seit 2012 arbeitet er in der Abteilung für Messtechnikentwicklung bei Berliner Glas an interferometrischen Messmethoden und rechnergestützter Datenauswertung. Zur Zeit konzentriert er sich auf die Verfeinerung der interferometrischen Messtechniken bei Berliner Glas.

Kontakt / *Contact*

✉ **richters@berlinerglas.de**

Nachdruck von / Reprint from:

Engelbert Hofbauer, Rolf Rascher, Manon Schilke, Johannes Liebl, Jan-Peter Richters (2016): Deflectometric acquisition of large optical surfaces (DaOS) using a new physical measurement principle: vignetting field stop procedure
Proceedings of SPIE Volume 10009: Third European Seminar on Precision Optics Manufacturing, 100090Y [Teisnach, April 12th 2016] doi:10.1117/12.2236134
Online verfügbar unter http://dx.doi.org/10.1117/12.2236134

DEC 2016 - NO. 2

BAVARIAN JOURNAL OF APPLIED SCIENCES®

Using Code Metrics for Android Programming

Peter Faber
Technische Hochschule Deggendorf

Stefan Schuster
Technische Hochschule Deggendorf

Tanja Maier
Technische Hochschule Deggendorf

ABSTRACT

Today, maintainability is of great importance for software projects. In this regard, software metrics play a crucial role in software development: these metrics may be used to objectively assess certain aspects of the software project at hand. We give an overview of available software metrics and evaluate their availability in software development tools. To that end, we explore their usage for the improvement of an Android app project – the E-WALD InCarApp. We provide evidence about their usefulness in a case study by measuring and comparing different aspects of the software project, leading to a derived software metric. We focus especially on measuring and improving code quality and compare these results to statements obtained from developer interviews which indicate that our derived metric may well be used to identify hot-spots for optimization.

Wartbarkeit ist heutzutage von größter Wichtigkeit für Software-Projekte. Hierzu spielen Software-Metriken eine zentrale Rolle in der Software-Entwicklung: Diese Metriken können genutzt werden, um gewisse Aspekte des betrachteten Software-Projekts objektiv einzuordnen. Wir geben eine Übersicht über zur Verfügung stehende Software-Metriken und evaluieren ihre Verfügbarkeit in Software-Entwicklungs-Tools. Dazu betrachten wir ihre Anwendung bei der Verbesserung eines Android-App-Projekts – der E-WALD InCarApp. Wir weisen ihre Nutzbarkeit in einer Fallstudie nach, in der wir unterschiedliche Aspekte des Software-Projekts messen und vergleichen, was uns zu einer abgeleiteten Software-Metrik führt. Wir konzentrieren uns hier vor allem auf die Messung und Verbesserung der Code-Qualität insbesondere der Wartbarkeit und vergleichen die Resultate mit Aussagen aus Interviews mit den Software-Entwicklern. Die einfache abgeleitete Metrik erscheint dabei durchaus schon geeignet, um Hot-Spots für Optimierungspotenziale zu identifizieren.

KEYWORDS

Java, Android, app programming, software engineering, software metrics

Java, Android, App-Programmierung, Software-Engineering, Software-Metriken

1. Introduction

Software development is a fast-paced process building systems in a world with constantly changing requirements. As a consequence, software source code evolves over time becoming more and more complex and thus harder to understand and maintain. To quantify this notion of complexity, many software metrics have been introduced and are available in different tools. These metrics can be used at different stages in the software development process, and each metric tends to show specific advantages and shortcomings. One could demand, for instance, that source files have to be short in order to be quickly understood and thus maintained. Such a demand could then be supported by counting the lines of code of all source files – this is the value the so-called Lines of Code metric (LOC) provides. However, even a short source file can be hard to modify if hundreds of classes depend upon this very file and one therefore also needs instruments measuring the degree of dependency.
Therefore, this work compares a range of popular code metrics and evaluates their fitness for refactoring purposes – in this case of the E-WALD Android InCarApp, an advanced driver assistance system for electric vehicles (EV). The selection of tools used for calculating metrics was therefore governed by their ability to process Android Java code. This article provides

- an introduction to the theory behind some of the more popular code metrics and gives an overview of the available tools to compute these metrics in Section 2.
- a case study describing the use of several metrics in the development process of the E-WALD InCarApp in Section 3. In particular, we provide measurements of different versions of the app and compare them.
- an evaluation of our case study and a discussion of how well the code metrics mirror the actual refactoring decisions in Section 4
- a conclusion and ideas for future work in Section 5.

2. Measuring Code Quality

Loosely following Ebert [1], most code metrics can be subsumed in one of the following categories:
* Volume-based metrics
* Encapsulation-based metrics
* Structure-based metrics
Let us first review the properties of these metrics following the above classification.

2.1 Volume-based Metrics

Volume-based metrics simply mirror the amount of code used for a given task -- just as the LOC metric mentioned in Section 1.

Number of Public Methods (NPM). Counts the number of public (and thus externally callable) methods of a class [2]. We use the tool ckjm [3] (see below) for measuring NPM.

Non-commented Source Statements (NCSS). Even though LOC is easy to compute, the lines that actually make up all the complexity of the code are only the executable source statements; thus, comments an empty lines should be ignored when comparing source codes [4]. NCSS only counts these executable source statements. This makes the metric harder to compute, since tools actually have to parse the code under consideration. Moreover, there is no common definition of „executable statement“. As an example, the Teamscale tool by Heinemann et al. [5] considers Java import statements as source code, while Tim Littlefair‘s CCCC [6] does not. This emphasizes that there is usually not a single absolute number that can be assigned to some piece of code even for simple metrics. We used the Teamscale tool to measure the NCSS metric.

Weighted Methods per Class (WMC). This metric, introduced by Chidamber and Kemerer [7], assigns a weight to each method of a class and sums up those weights. This gives a rough estimate of the complexity of the class. By adjusting, the weights, can be adapted to specific situations. The most common weights are:

1. simply assign 1 to each function (counting the number of functions)

2. compute McCabe‘s cyclomatic complexity for each function (see below).

WMC -- along with a whole set of additional metrics -- can be computed using the open source command line tool ckjm [3] (which uses weight 1) or SourceMeter [8] (which uses McCabe‘s cyclomatic complexity). We use ckjm for our measurements, since we view McCabe‘s complexity separately.

McCabe‘s Cyclomatic Complexity (McCabe). The intuition behind McCabe‘s cyclomatic complexity is to measure the number of decisions in a function. To be more precise, it represents the number of independent paths (without any assumptions on decision results) through the control flow graph (CFG) of a function. The cyclomatic complexity C(G) is defined through the CFG G=(V,E) of the function [9, 10]. For a given piece of code, let V be the set of its code blocks (basic blocks). We connect two code blocks by an edge, if one code block can be directly executed after the other. Let E then be the set of those edges. Then, the cyclomatic complexity C(G) is defined by:

(1) $$C(G) = |E| - |V| + 2$$

As an example, Figure 1 shows a CFG for the code if (B){ X; }. In this case, |V| = 4, |E| = 4, resulting in an overall complexity of C(G)= 4 - 4 +2 = 2, which is in fact the number of possible paths of execution. Thus, McCabe can be used to estimate the number of test cases needed for complete branch coverage. It also provides a hint towards maintainability (due to software complexity), since a code containing many complex decisions is harder to understand and thus harder to maintain than a simple one-dimensional sequence of instructions.

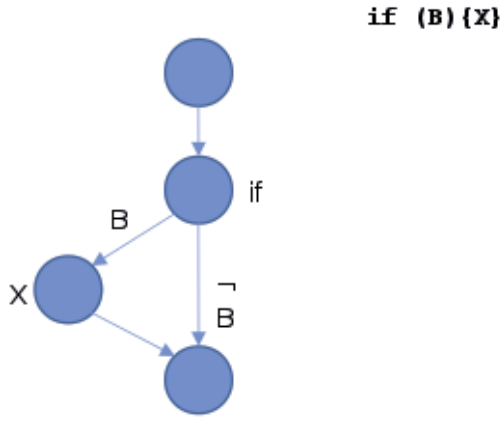


Figure 1: Example control flow graph for if (B){ X; }.

Tools that measure McCabe‘s cyclomatic complexity include CCCC and SourceMeter [8, 11]. Again, there are differences in the definitions applied by the tools. The most important difference here is whether short-circuiting in the programming language (i.e., evaluating a Boolean expression only as long the outcome of the expression is not yet clear, e.g. in true or [...]) is assumed to create paths in the CFG or not. For our experiments, we use SourceMeter, which models short-circuits as branches in the CFG.

(Maximum) Nesting Depth (MND), also called Nested Block Depth (NBD). The maximum nesting depth is defined as the maximum number of control statements nested into each other [12]. We measure nesting depth using the Teamscale tool.

2.2 Encapsulation-based Metrics

Object-oriented languages like Java pose a challenge for volume-based metrics:

* The methods in classes of object-oriented languages tend to be quite short and to hide decision complexity -- as the one measured by McCabe -- by calling other methods.

* Moreover, object-oriented programming itself introduces a completely different kind of complexity in programs that volume-based metrics do not handle at all: relations between classes and packages.

To tackle these problems, several new metrics have been introduced in the literature.

Coupling Between Object classes (CBO). This metric counts the number of classes that

a certain class A communicates with (i.e. the number of classes B from which A calls methods or references variables) [7]. If A communicates with B on several occasions, this relation is still only counted once. Chidamber and Kemerer [7] cite 3 reasons for introducing this metric:
* A high coupling degree reduces modularity, and reuse is hindered.
* A small coupling degree promotes encapsulation and thus improves maintainability.
* Higher inter-object class coupling entails the need for more rigorous tests (and thus more complex tests).
Some authors exclude those classes B that are already in an inheritance relation with A [1]. To compute this metric, we use ckjm, which employs the original definition by Chidamber and Kemerer [7].

Afferent Couplings (CA) / Efferent Couplings (CE). Besides the mere number of coupled classes as measured by the CBO metric, one may also be interested in the strength of this coupling. Afferent (incoming) Couplings CA(A) of a class A is defined as the number of members of A that are accessed by some other class B. Correspondingly, Efferent Couplings represent the sum of all members defined in some class B that class A uses.
As an example, consider Listing 1.

```
class B{
    public void doSomething(){...}
    public void doSomethingElse(){...}
}
class A{
  B b;
    public A() {
        B b = new B();
        b.doSomething();
        // re-inserting the following statement
        // increases RFC(A) from 4 to 5:
        // b.doSomethingElse();
    }
    private void doSomething() {
        b.doSomething();
    }
}
```

Listing 1. Example Java program: class B depends on class A via calls to doSomething() and doSomethingElse().

In the example of Listing 1, class A calls methods of class B in several places. CBO(A)=1, since only class B is referenced. Moreover, since A only calls B.doSomething() (and no additional function) CE(A)=CA(B)=1. This holds, although B.doSomething() is called twice. However, if the call to B.doSomethingElse() in A.doSomething() is re-inserted, A now calls two different functions from B, increasing CE(A)=CA(B) to 2.

Depending on application and author, the definition of CA/CE can differ, e.g. to mean the number of classes outside the current package that refer to a class inside the current package [12].
We calculate only CA and use ckjm to that end.

Response For a Class (RFC). Just as CBO, RFC was originally introduced by Chidamber and Kemerer [7]. Essentially, RFC(A) is the same as CE(A) plus the number of methods defined in A itself. Formally, RFC for a class A is defined as the number of different methods (defined in some other classes) that are called by methods of A plus the number of methods defined in A.
Reconsider the code snippet from Listing 1. In this case, RFC(A)=2+1+1=4: A contains 2 methods (the constructor B.B(), and B.doSomethingElse()), it calls 1 method from class A (A.doSomething()), and, since in Java, all objects inherit from class Object, it also calls the constructor of Object as its parent class.
Note again that, although RFC(A) does not increase any further due to the repeated call to B.doSomethingElse(), it does increase to 5, if the call to B.doSomethingElse() that is commented out is re-inserted into the code, since a different method of B is now called in addition to B.doSomethingElse().
Thus, the response for A is the number of methods that can theoretically be called when a message is sent to an object objA of type A by some other object objB (i.e. when a method of objA is called).
Again, Chidamber and Kemerer [7] cite 3 reasons for introducing this metric:
* Testing and debugging may become increasingly complex when a large number of methods has to be executed (and thus analyzed during debugging) as a response to a received message.
* Thus, RFC may be used to assess testing time.
* A large number of member methods may hint at an altogether complex class.
We measured this metric using ckjm.

2.3 Structure-based Metrics

In addition to volume-based and encapsulation-based metrics, there are further metrics that can be helpful in the analysis of source code. We subsume these metrics under the general term of structure-based metrics.

One important point – the only one we consider in this work – is the existence of clones. Clones

are exact copies of code that occur at different sites from the original code. For example, the call b.doSomething() in Listing 1 occurs twice in class A -- once in the constructor and once in method A.doSomething(). Both of these occurrences are clones of each other. However, as one can plainly see, these small clones of length 1 (line) can even be desirable. However, larger sections of code that occur several times in some program code may hint at unstructured source code, because clones usually could very well be re-combined into a single function that can be called from several places. Therefore, it is important to choose a threshold that defines a lower bound for the size of a clone in order to be considered a clone. Clone detection refers to the pure indication of recurrences which can be quite helpful during coding itself. Based on the number and sizes of clones, a number can be computed to give the clone sites a meaning as a metric. This leads us to clone coverage.

Clone Coverage (CC). The canonical metric for clone recognition as described above is Clone Coverage, which is defined as the percentage of source code detected to be a clone. This can be interpreted as the probability with which some randomly picked source code is actually (part of) a clone [13]. Parameters influencing clone coverage include not only the minimal length of a clone (in LOC), but also whether variable names should be unified (or interpreted as-is) and whether generated code is to be excluded. We used the Teamscale tool to compute CC. In our analysis, clones have a minimal length of ten lines, third-party libraries were excluded (which also holds for the other metrics applied here), and variable names are subjected to a unification algorithm, i.e. clones can be recognized even after (a simple) renaming.

3. Tools

This section gives a brief overview of some tools that can be used to compute the previously defined metrics. We will describe each tool briefly and give a short assessment of the usability of the tool.

ckjm. Short for Chidamber and Kemerer Java Metrics, ckjm is a tool developed by Spinellis [3] to compute the metrics by Chidamber and Kemerer for Java programs [7]. In their paper, Chidamber and Kemerer propose seven metrics for software analysis: WMC (weighted methods per class), CBO (coupling between object classes), RFC (response for a class) plus three more metrics that we did not discuss above – DIT (depth of inheritance tree), NOC (number of children), and LCOM (lack of cohesion in methods). These metrics are computed directly as defined in the original paper by Chidamber and Kemerer [7]. In addition, Spinellis [2] included CA (afferent couplings) and NPM (number of public methods) in his tool. The tool is a stand-alone console program. The computed results are presented directly on the standard output.

ckjm computes metric values for each class separately. Anonymous classes are handled as autonomous entities. For the analysis of Java programs, ckjm relies on byte code (.class) files in JVM format. This imposes a technical problem for Android projects, where the target is Dalvik format (Dex byte code). However, Java classes are usually first compiled into JVM format before the conversion into Dex byte code. So it is possible to use the intermediate JVM byte code files (whose location depends on the IDE used) for analysis through ckjm. Also, input files have to be provided correctly on the command line. This may include additional work with UNIX tools such as find or xargs. However, several files may be combined into a single report. Still, there is no plug-in available to directly view the results during coding in an IDE.

ckjm is an open source Java program and claims to run on any Java-enabled platform.

Teamscale. The software and consulting company Continuous Quality in Software Engineering (CQSE) develops a GUI-based tool called Teamscale which is also available in an open source version called ConQAT [5, 17]. Teamscale is designed to support analysis, monitoring, optimization, and code quality engineering during software development. Teamscale shows its results in real-time.

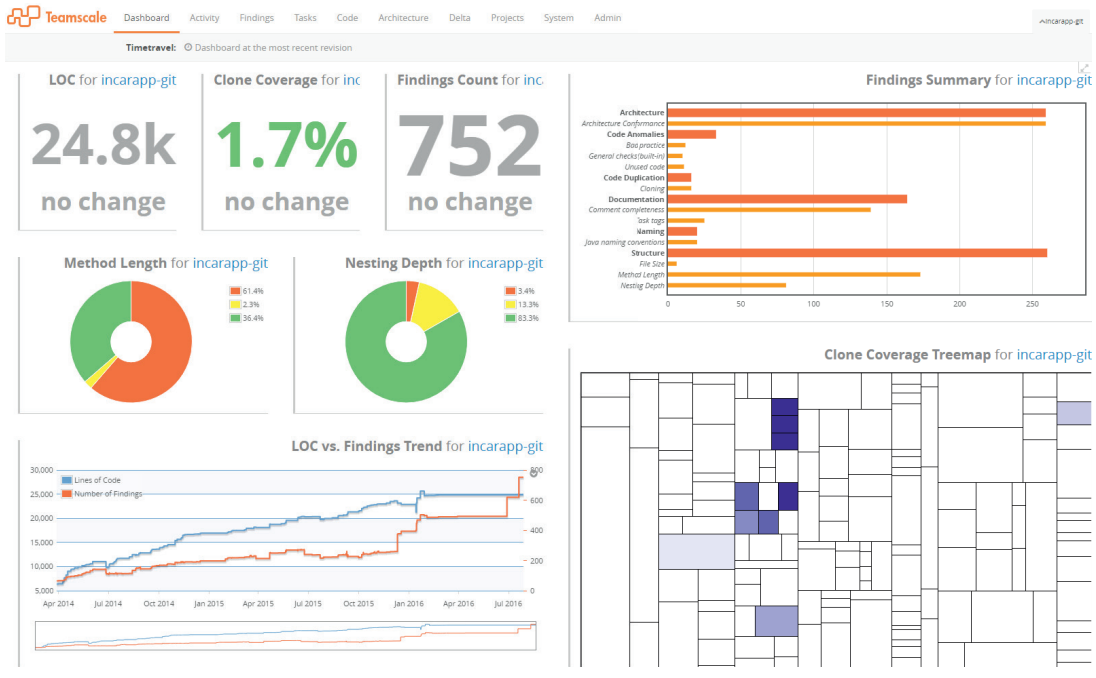


Figure 2. Dashboard of the Teamscale GUI.

The main features of the Teamscale analysis tool include analysis for architecture conformance, clone detection, test gap analysis, coding conventions, and documentation analysis.
Teamscale consists of an analyzing backend that can be run on a development computer or on a server, and a frontend that is used to represent Teamscale's findings graphically; the frontend can be run as a stand-alone GUI-based program or from within an IDE as plug-in (currently available for Eclipse, Visual Studio and IDEA). The most important metrics are directly visible on a configurable dashboard (cf. Figure 2).
This approach is very well suited for immediate feedback to the developer and interactive code reviews: The values observed can be represented on a time line, and the representation as a tree map directly shows very clearly where thresholds for certain metrics are exceeded in a package by applying a color code to areas of a map.
However, the tool appears suboptimal for the generation of offline reports, since the GUI partly relies on interactivity (such as the mouse hovering over interesting parts of a diagram) in order to reveal exact numbers from the analysis.
Teamscale and ConQAT are Java programs; the server component of Teamscale is available for Linux and Windows.

CCCC. CCCC is another text console-based stand-alone program. The current Ubuntu-based Linux distributions even contains a CCCC package. For each source file, it creates an XML and an HTML file that represent the results. The files are located in a hidden directory called .cccc. CCCC computes the metrics of Chidamber and Kemerer. Additionally, it computes some volume-based metrics such as LOC and McCabe. Originally, CCCC (short for C/C++ code counter) was developed for C/C++ code. CCCC claims to be also able to parse Java code, so that it can be used in the context of Android app programming.
Indeed, CCCC was able to produce output for the complete E-WALD InCarApp. However, not all Java code was treated correctly by this tool: Java annotations (such as @override) seem to confuse CCCC so much that it skips the corresponding function entirely and assigns 0-values. Since CCCC produces two output files for each input file, additional work has to be done to extract information for a common report; however, to that end, the XML files produced can be converted using UNIX tools. Thus, CCCC can be used quite well to generate offline reports using scripts; however, it does not deliver immediate feedback via IDE plug-ins.
CCCC is an open source program available for Linux, FreeBSD, and Windows.

SourceMeter. Another console-based tool is SourceMeter [8, 11], which is a program developed at FronteEndART, a company

specializing in software quality management. It can be used for Java, C/C++, C#, Python, and RPG. SourceMeter is able to compute a plethora of metrics (approx. 70), e.g. McCabe, WMC, CBO, RFC, depth of inheritance tree (DIT), LOC, number of attributes (NA), and number of classes (NCL). The metric definitions here differ slightly from those employed by ckjm; e.g., as already mentioned, McCabe is used as weight for each function in WMC. SourceMeter is well-adjustable in parameters for the computed metrics (such as minimal clone length) and supports more input languages than usual.
SourceMeter is supported on Windows and Linux.

Sonargraph. Sonargraph is both, a plug-in for Eclipse/Spring or IDEA, and a GUI-based tool for analyzing Java, C#, and C++ code. It is available in two flavors -- the free Explorer version and the commercial Architect version, which includes more features. Sonargraph specializes on a specific set of metrics that we did not consider in this first study [14]. These metrics are based on class dependencies as observed by Robert C. Martin [15]; the basic idea here is that cycles in the dependency graph of classes should be avoided. However, in our preliminary tests, these kinds of dependencies occurred mainly due to the Android operating system.
Sonargraph is available for Windows, Linux, and MacOS.

Checkstyle. As an aside, this tool should be mentioned as a valuable helper for Java development; it is not directly used to measure metrics, but it can be used to guard coding standards that have been set in a project. Checkstyle, as its project homepage says, is a development tool to help programmers write Java code that adheres to a coding standard [16]. This tool can be executed as an Ant task, but also as a plug-in for Eclipse, IDEA or NetBeans. If used as a plug-in, it directly marks the occurrences of code segments that do not adhere to the defined standards. These coding standards can be adjusted by defining patterns and thresholds for these patterns. Checkstyle supplies immediate feedback to the programmer and may therefore be quite helpful in producing standard conforming code from the start.
Checkstyle is an open source stand-alone tool that can also be called via plug-ins for several IDEs including Eclipse and claims to run on Java platforms. Current Ubuntu-based Linux distributions include Checkstyle.

Summary. Table 1 shows a comparison matrix for the different tools. The first three columns show whether the corresponding tool can be used as plug-in in an IDE, in a text-console, or with a GUI. The following 6 columns show the most important supported programming languages for each tool, and the last set of columns indicates which of the more common metrics can be measured using the corresponding tool.

	UI			Supported programming languages						Metric									
Name	Plug-In	Console	GUI	Java	C/C++	C#	Python	JavaScript	ABAP	LOC	NCSS	WMC	NPM	McCabe	CBO	RFC	CA	CR	CC
ckjm - Chidamber and Kemerer Java Metrics	✗	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	✓	✓	✓	✗	✗
Teamscale/ConQAT	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✓	✓
Checkstyle	✓	✓	✗	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗	✓	✓	✗	✗	✓	✗
Sonargraph Architect	✓	✗	✓	✓	✓	✓	✗	✗	✗	✗	✓	✗	✗	✓	✗	✗	✗	✓	✗
CCCC	✗	✓	✗	✓	✓	✗	✗	✗	✗	✗	✓	✓	✗	✓	✓	✗	✓	✗	✗
SourceMeter	✗	✓	✗	✓	✓	✓	✓	✗	✗	✓	✗	✓	✓	✓	✓	✓	✗	✓	✗

Table 1. Overview of software metrics tools and their features.

4. Evaluation

In order to evaluate the use of software metrics in the development process and assess the usefulness of the tools above, the authors followed the development process of the E-WALD InCarApp.

The E-WALD InCarApp represents a vital point in the E-WALD project whose aim is to support electric car concepts available today in rural areas. The InCarApp itself is an Android app installed on tablets inside the E-WALD electric vehicles. It is used for (1) collecting data about the current state of the car of a trip, and (2) for updating the driver about charging stations, the expected remaining range of the car on a map, and other information.

(1) Regarding the first of those aspects, the app‘s purpose can be divided into two phases that build a kind of data pipeline:

1. Collect data from
* the car‘s CAN bus (CAN: Controller Area Network; a car-internal communication network),
* GPS coordinates,
* tablet data (temperature, battery state of health etc.),
* possibly further data.
2. Provide the collected data via a central hub, e.g. for computing the remaining range of the EV.
3. Process collected data in different ways, e.g.:
* send data packets to a server for further storage and data processing
* internally process data for visual user feedback in the car

Separating these tasks from each other in the logic is of importance for the reusability of the corresponding code fragments. Thus, in order to reuse the InCarApp code in other projects, this structure should be mirrored in the class structure.

The InCarApp is in the process of transitioning from version 11 to version 12. Version 11 of the InCarApp does not clearly separate the three aspects identified, as can be gleaned from Figure 3, which shows part of the original InCarApp in a class diagram.

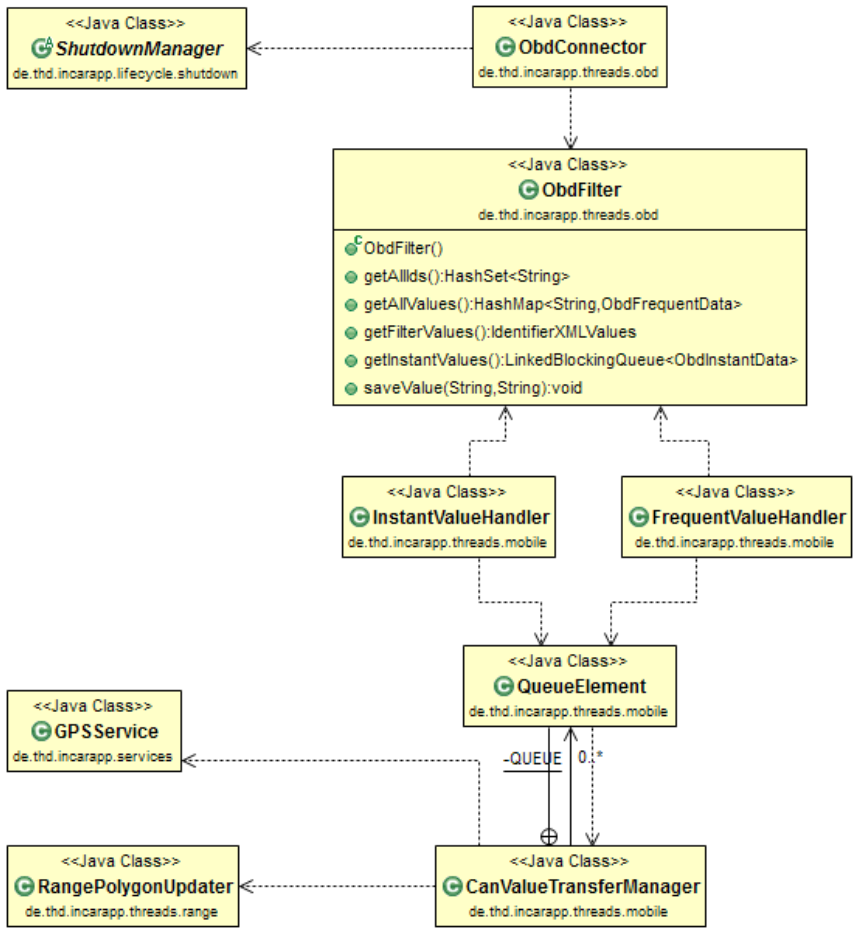


Figure 3. E-WALD InCarApp class structure, version 11.

A crucial point here that lends itself to optimization can be seen in the class CanValueTransferManager. This is actually the main data hub of the original app and is responsible for sending data to a server. However, this class was originally designed to send CAN values only (as can already be guessed from its name) and is therefore not well extensible to include other data as well. In fact, in order to send additional data such as GPS coordinates, it is necessary to modify the code of CanValueTransferManager (it has to pull the new value from the class GPSService). Instead, one would rather like to simply submit new data via some connection broker.
In addition, there are direct dependencies to classes which are responsible for rather unrelated aspects of the overall app. For instance, in case of a lost connection to the CAN bus, the class ObdConnector directly tries to restart the app by sending a message to ShutdownManager. In another case, CanValueTransferManager also decides that a new computation should be triggered in class RangePolygonUpdater, although this decision has nothing to do with the communication itself.
Also, the frequency in which data is sent to the server has to be made explicit in the corresponding classes. This can also be seen in the relationship between classes.
Large parts of the original InCarApp have been refactored in an attempt to achieve those goals. The current version 12 encapsulates each of these aspects into stand-alone libraries which are completely independent from each other. The app was divided into two major parts during refactoring: on one hand, a helper project — the CAN-lib — was created that contains all necessary functionality to talk to the CAN bus, on the other hand, the remaining InCarApp contains the central data hub, but also functionality for server communication and end user presentation. For the purpose of this study, these parts where always viewed together as a single application.
Both the earlier version 11 and the refactored version 12 of the E-WALD InCarApp where subjected to a series of analyses using different metrics. In the following, we report about the result of this examination.
In this work, we report on the results of following metrics:
* Weighted Methods per Class (weight 1)
* Number of Public Methods
* Non-Commented Source Statements
* McCabe‘s Cyclomatic Complexity
* Maximum Nesting Depth
* Coupling Between Object classes
* Afferent Coupling
* Clone Coverage (minimum clone length: 10 lines)

For each metric M, we calculate the maximum value observed across the different classes C:

$$Max_M = max_{C \in classes} M(C) \quad (2)$$

For each class C and each metric M, we then build the ratio

$$\frac{M(C)}{Max_M} \quad (3)$$

The rank of a class C, R(C), then is given by the sum across all metrics considered:

$$R(C) = \sum_{M \in metrics} \frac{M(C)}{Max_M} \quad (4)$$

Thus, the highest rank specifies the class with the highest metric values overall. In our case, for each metric, a higher value means a less favorable outcome (our aim are few, concise, uncoupled methods and classes).

	Volume					Encapsulation		Structure	
	WMC	NPM	NCSS	McCabe	MND	CBO	CA	CC (%)	R(C)
ChargingMapActivity	122 (100%)	65 (100%)	1165 (100%)	91 (100%)	5 (71%)	199 (100%)	50 (37%)	0,0 (0%)	6,08
FileServerThread	29 (24%)	15 (23%)	547 (47%)	57 (63%)	4 (57%)	32 (16%)	6 (4%)	0,0 (0%)	2,34
SettingsActivity	48 (39%)	33 (51%)	431 (37%)	29 (32%)	2 (29%)	66 (33%)	13 (10%)	0,0 (0%)	2,30
DebugErrorTab	21 (17%)	8 (12%)	290 (25%)	35 (38%)	7 (100%)	35 (18%)	7 (5%)	0,0 (0%)	2,16
Logger	16 (13%)	11 (17%)	163 (14%)	17 (19%)	2 (29%)	9 (5%)	136 (100%)	0,0 (0%)	1,96
CanValueTransferManager	22 (18%)	13 (20%)	418 (36%)	43 (47%)	4 (57%)	16 (8%)	6 (4%)	0,0 (0%)	1,91
StartupLogoActivity	33 (27%)	19 (29%)	195 (17%)	12 (13%)	5 (71%)	43 (22%)	14 (10%)	0,0 (0%)	1,90
HomeButtonActivity	28 (23%)	12 (18%)	251 (22%)	15 (16%)	6 (86%)	36 (18%)	8 (6%)	0,0 (0%)	1,89
PoiUpdater	15 (12%)	8 (12%)	330 (28%)	54 (59%)	4 (57%)	21 (11%)	7 (5%)	0,0 (0%)	1,85
PoiGroupGridViewAdapter	22 (18%)	15 (23%)	288 (25%)	35 (38%)	4 (57%)	23 (12%)	4 (3%)	0,0 (0%)	1,76

Table 2. Top ten worst rated classes of InCarApp version 11.

Table 2 shows the top ten classes in descending order according to their rank R(C) as they appear in version 11 of the E-WALD app. The whole app contains 151 classes. Metric values here are supplied as absolute numbers, but also as percentage with respect to the corresponding maximum values in the project. Table 1 shows ChargingMapActivity as the number one target for refactoring. CanValueTransferManager, a class that had already been speculated as a good candidate for refactoring, ranks at number six among the top ten highest ranking classes. Further classes in this top ten list include FileServerThread, PoiUpdater, which had also already been on the developers' list. In contrast to those classes, HomeButtonActivity, number eight on the top ten list, had not been an a-priori candidate for the developers. However, a closer examination reveals that the class contains several unnecessarily deeply nested try blocks. These make the implementation harder to understand and were flagged due to the high value of MND. On the other hand, there are classes like Logger that rank quite high in the list, although they cannot be identified as a valid target for refactoring. In the example of the Logger class, this is primarily due to the fact that the class has to be called from almost all other classes, and is therefore coupled quite closely to those other classes (yielding a CA rating of 136, which is the maximum CA value of all classes).

	Volume					Encapsulation		Structure	
	WMC	NPM	SLOC	McCabe	MND	CBO	CA	CC (%)	R(C)
MapActivity	50 (100%)	17 (52%)	417 (100%)	43 (96%)	3 (75%)	101 (100%)	15 (63%)	0,0 (0%)	5,85
ReserveChargingStation	50 (100%)	10 (30%)	238 (57%)	18 (40%)	4 (100%)	66 (65%)	12 (50%)	0,0 (0%)	4,43
DataSenderService	35 (70%)	18 (55%)	310 (74%)	30 (46%)	3 (75%)	45 (45%)	12 (50%)	20,2 (20%)	4,35
EwaldMap	26 (52%)	17 (52%)	221 (53%)	39 (87%)	4 (100%)	41 (41%)	7 (29%)	0,0 (0%)	4,13
DataSenderService	23 (46%)	13 (39%)	197 (47%)	22 (49%)	3 (75%)	28 (28%)	8 (33%)	0,0 (0%)	3,18
FakeCanConnectionV1	30 (60%)	11 (33%)	229 (55%)	21 (47%)	2 (50%)	9 (9%)	7 (29%)	0,0 (0%)	2,83
CarLocationOverlay	19 (38%)	10 (30%)	174 (42%)	23 (51%)	3 (75%)	27 (27%)	2 (8%)	0,0 (0%)	2,71
CanService	22 (44%)	12 (36%)	122 (29%)	7 (16%)	2 (50%)	39 (39%)	13 (54%)	38,9 (39%)	2,68
OnNavigationIconHeaderClickListener	23 (46%)	6 (18%)	106 (25%)	9 (20%)	4 (100%)	27 (27%)	7 (29%)	0,0 (0%)	2,66
EwaldChargingApiReader	11 (22%)	2 (6%)	174 (42%)	45 (100%)	3 (75%)	7 (7%)	1 (4%)	0,0 (0%)	2,56

Table 3. Top ten worst rated classes of InCarApp version 12.

Table 3 shows the same top ten ranking for the 205 classes of version 12 of the E-WALD InCarApp. MapActivity (which had been refactored into MapActivity and the also high-ranking EwaldMap) still leads the ranking of classes to be refactored. Nevertheless, although this class remains a candidate for further improvement, the absolute values were remarkably reduced for all metrics considered, indicating that the refactoring effort indeed showed some effect. Again, the classes that leave room for improvement in the view of the developers also rank high according to our top ten list. An example is DataSenderService, ranking on place five. This class is still thought to have too high a degree of different responsibilities, which could be improved upon in further releases.

In order to evaluate the changes between version 11 and 12 of the E-WALD InCarApp, we also considered the average rating of all classes according to the different metrics considered.

For each metric M, we calculated the average value A(M) of the metric across all classes of an app version:

$$A(M) = \frac{\sum_{C \in classes} M(C)}{number\ of\ classes} \quad (5)$$

Figure 4 shows the ratio of these averages between version 11 and 12 of the InCarApp.

As can be gleaned from Figure 4, almost all relevant metric values have been reduced on average in version 12 of the app, with the lowest reduction in the number of methods (WMC): the new version still uses 79% of the number of methods per class in comparison to the previous version. The highest reduction can be seen in Afferent Couplings (CA), which now average at only 50% of the previous values.

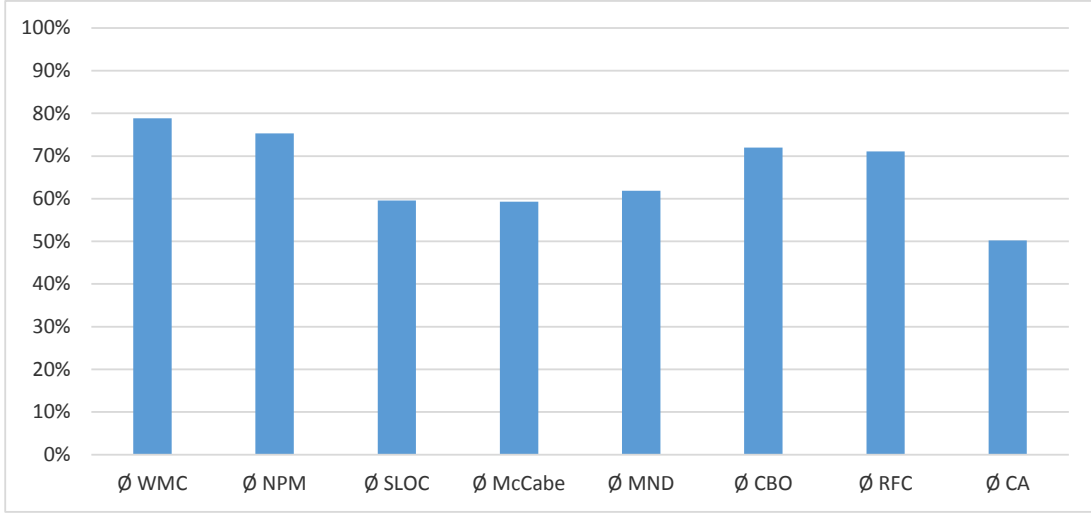


Figure 4. Relative changes between version 11 and version 12 of the InCarApp; the metric values for version 11 correspond to 100%.

Still, there is one exception that is not explicitly shown in Figure 4: The clone coverage actually increased from 1.7% to 8.7% in the new version. This is probably due to the fact that the InCarApp is still in the process of refactoring, in which certain classes have been copied in order to guarantee functional equivalence during recoding.

5. Conclusion and Future Work

Overall, our findings can be summarized as follows:

* Software metrics can indeed hint quite well at hot spots in a software project, even in the case of a larger, grown Android app project.
* Nevertheless, the decision about which hot spots are really in need of a redesign should never be made blindly following the numbers any code metric (or combination thereof) produces (e.g. a logger class will usually be coupled with many other classes).
* Using a ranking method like mentioned above may not yield an authoritative answer as to which packages and classes should be refactored, but it does give a good priority list on which classes to have a look at first.
* For agile development methods in particular – that generally work with frequent code changes – plug-ins can greatly improve development. However, encapsulation-based metrics – which could help in the coding process – are almost never found in tools that can be used as plug-ins. The metrics considered here are strictly syntax-based. An interesting point lies in how far semantic aspects could be integrated into code metrics, e.g. simply by identifying semantically related expressions or defining structural patterns. As a first step, however, more metrics from different fields should be included in our consideration, and the crucial point of creating a ranking function – which may be quite project-specific – requires thorough quantitative analysis.

6. Disclaimer

The authors are not aware of any conflicts of interest.

7. Acknowledgements

The authors wish to thank the software developers of the E-WALD project team, in particular Mr. Michael Schönberger, who supported this work in numerous interviews and discussions.
The E-WALD project is funded by the Bavarian State Ministry for Economic Affairs and Media, Energy and Technology.

8. References

[1] Ebert, Christof; Dumke, Reiner (1996): *Software-Metriken in der Praxis. Einführung und Anwendung von Software-Metriken in der industriellen Praxis.* Berlin et al.: Springer.

[2] Spinellis, Diomidis (2005): "Metric Descriptions." [Online]. Available: http://www.spinellis.gr/sw/ckjm/doc/metric.html. (Accessed: June 20, 2016).

[3] Spinellis, Diomidis (2005): "Tool Writing: A Forgotten Art?" In: IEEE Software: IEEE Computer Society Press, Los Alamitos, CA, USA 22 (4), pp. 9–11.

[4] Schneider, Kurt (2012): *Abenteuer Softwarequalität. Grundlagen und Verfahren für Qualitätssicherung und Qualitätsmanagement.* Heidelberg: Dpunkt.verlag.

[5] Heinemann, Lars; Hummel, Benjamin; Steidl, Daniela (2014): "Teamscale: Software Quality Control In Real-Time." In: Companion Proceedings of the 36th International Conference on Software Engineering (ICSE ,14). New York: ACM, pp. 592–595.

[6] Littlefair, Tim (2001): An Investigation into the Use of Software Code Metrics in the Industrial Software Development Environment. Ph.D. Thesis. Edith Cowen University, Perth, Western Australia. Faculty of Communications, Health and Science.

[7] Chidamber, Shyam R.; Kemerer, Chris F. (1994): „A Metrics Suite for Object Oriented Design." In: IEEE Transactions on Software Engineering 20 (6), pp. 476–493.

[8] Ferenc, Rudolf; Langó, László; Siket, István; Gyimóthy, Tibor; Bakota, Tibor (2014): "Source Meter Sonar Qube Plug-in." In: Proceedings of the 14th IEEE International Working Conference on Source Code Analysis and Manipulation (SCAM) 2014, pp. 77–82.

[9] McCabe, Thomas J. (1976): "A Complexity Measure. In: Proceedings of the 2nd International Conference on Software" Engineering (ICSE ,76): IEEE Transactions on Software Engineering (Vol. SE-2, 4), pp. 308–320.

[10] Liggesmeyer, Peter (2009): S*oftware-Qualität. Testen, Analysieren und Verifizieren von Software.* 2nd Ed. Heidelberg: Spektrum Akademischer Verlag.

[11] FrontEndART: SourceMeter Project Homepage. [Online]. Available: https://www.sourcemeter.com. (Accessed: July 24, 2016).

[12] Oliveira, Marcio F.S.; Redin, Ricardo Miotto; Carro, Luigi; da Cunha Lamb, Luís; Wagner, Flávio Rech (2008): "Software Quality Metrics And Their Impact On Embedded Software." In: 5th International Workshop on Model-based Methodologies for Pervasive and Embedded Software (MOMPES) 2008: IEEE, pp. 68–77.

[13] Göde, Nils; Hummel, Benjamin; Juergens, Elmar (2012): "What Clone Coverage Can Tell." In: Proceedings of the 6th International Workshop on Software Clones (IWSC ,12) (IEEE Computer Society), pp. 90–91.

[14] hello2morrow GmbH: Sonargraph Homepage. [Online]. Available: https://www.hello2morrow.com/doc/sg7/index.html. 2014. (Accessed: July 30, 2016).

[15] Martin, Robert C. (1995): Object Oriented Design Quality Metrics: An Analysis Of Dependencies. In: C++ Report (SIGS Publications Group) Software Metrics and Android Programming (September/October).

[16] Checkstyle Project: Checkstyle Homepage: [Online]. Available: http://checkstyle.sourceforge.net/index.html (Accessed: July 29, 2016).

[17] CQSE: Teamscale Homepage. [Online]. Available: https://www.cqse.eu/en/products/teamscale/features. (Accessed: July 27, 2016).

Prof. Dr. Peter Faber

Peter Faber holds a Diploma degree in Computer Science from the University of Passau. From the same institution he received his doctorate on the subject of code optimization in the polyhedron model. Between 1999 and 2004, he worked as a scientist at GMD – Forschungszentrum Informationstechnik – and at the University of Passau, and from 2005 to 2009 as IT consultant and software engineer at science+computing ag. Since 2009, he has been professor for Media Technology at Technische Hochschule Deggendorf (THD). Since 2015, he has been the academic head of THD's Computing Center. In research and teaching, his activities include programming and code optimization, in particular graphics and parallel programming, as well as programming of mobile devices.

Peter Faber erhielt sein Diplom in Informatik von der Universität Passau. Dort promovierte er auch zum Thema der Optimierung von Schleifenprogrammen im Polyedermodell (Code optimization in the polyhedron model).
Zwischen 1999 und 2004 arbeitete er als Wissenschaftler an der GMD – Forschungszentrum Informationstechnik – und an der Universität Passau und von 2005 bis 2009 als IT-Consultant und Software-Engineer bei der science+computing ag. Seit 2009 ist er Professor für Medientechnik an der Technischen Hochschule Deggendorf (THD), seit 2015 zudem wissenschaftlicher Leiter des Rechenzentrums der THD. In Forschung und Lehre ist er aktiv im Bereich der Programmierung und Codeoptimierung, insbesondere der Grafik- und Parallelprogrammierung (GPGPU) und der Programmierung mobiler Geräte.

Kontakt / *Contact*
peter.faber@th-deg.de

Tanja Maier, B. Eng.

Tanja Maier is a software engineer at Innowerk-IT GmbH, a software and IT company. She received her Bachelor of Engineering degree in Media Technology with a focus on Media Computer Science from Technische Hochschule Deggendorf (THD). She subsequently enrolled in THD's Master's Program Media Technology with an emphasis on Industrial Multimedia. As a student trainee, she has been developing software within various software projects while studying at THD. For her Master's thesis, she is examining the role of software metrics in practical experience.

Tanja Maier ist Software-Engineer bei Innowerk-IT GmbH, einem Software- und IT-Beratungsunternehmen. An das Studium der Medientechnik an der Technischen Hochschule Deggendorf (THD) mit Schwerpunkt Medieninformatik (B. Eng.) schloss sie das Masterstudium Medientechnik und -produktion an der THD mit Schwerpunkt Industrielles Multimedia an. Neben dem Studium arbeitete sie bereits als Werkstudentin in verschiedenen Softwareprojekten als Software-Entwicklerin. Im Rahmen ihrer Master-Arbeit betrachtet sie die Rolle von Software-Metriken in der Praxis.

Kontakt / *Contact*

✉ **tanja.maier@innowerk-it.de**

Dipl.-Inf. Stefan Schuster

Stefan Schuster received his Diploma degree in Computer Science from the University of Passau, where he worked on several topics related to energy generation, such as the Algebraic Oil Project. Within this project, he developed symbolic-numeric methods to analyze oil production data. Since December 2014, he has been the responsible for the development of the Optimized Range Model within the E-WALD project, the largest demonstration project for electromobility in Germany.

Sein Informatikstudium absolvierte Stefan Schuster an der Universität Passau. Als wissenschaftlicher Mitarbeiter arbeitete er dort von 2008 bis 2014 in verschiedenen Projekten an Themen aus dem Bereich Energiegewinnung. So entwickelte er beispielsweise im Rahmen des Algebraic Oil Projects symbolisch-numerische Methoden zur Analyse von Ölproduktionsdaten. Seit Dezember 2014 ist er im E-WALD-Projekt u.a. für die Umsetzung des Optimierten Reichweitenmodells verantwortlich.

Kontakt / *Contact*

✉ **stefan.schuster@th-deg.de**

BAVARIAN JOURNAL OF APPLIED SCIENCES®

Exchange Rate Pass-Through into Swiss Prices[1]

Kersten Kellermann
Technische Hochschule Deggendorf

Carsten-Henning Schlag
Konjunkturforschungsstelle Vierländereck (KOVL)

ABSTRACT

The effective exchange rate of the Swiss franc has soared since 2007. Yet the appreciation has led only to a minor adjustment of Swiss import prices, consumer prices for imported goods and overall consumer prices. We apply an approach proposed by Campa and Goldberg (2005) to show that exchange rate pass-through (ERPT) is generally low, but slightly higher with respect to import prices than with respect to consumer prices. It increases when the adjustment period is extended. Our estimation results show that the exchange rate changes are largely absorbed in cross-border trade. In the time period of 1980-2007 Swiss consumers of imported goods felt only around 27% of the initial exchange rate shocks.

JEL Classification: E31; F31; F41; C22

Der Außenwert des Schweizer Frankens ist seit 2007 stark angestiegen. Die Aufwertung hat jedoch nur zu einer relativ geringen Anpassung der Schweizer Importpreise, Konsumentenpreise für importierte Güter sowie des Konsumgüterpreisindexes geführt. Im Beitrag wird – basierend auf einem Ansatz von Campa und Goldberg (2005) – gezeigt, dass die Wechselkurselastizität Schweizer Preise generell gering ist. Die Übertragungseffekte (exchange rate pass-through, ERPT) im grenzüberschreitenden Handel sind dabei etwas höher als die Übertragungseffekte in Bezug auf die Konsumentenpreise. Die Reagibilität erhöht sich, wenn die Anpassungsphase verlängert wird. Unsere Schätzergebnisse legen den Schluss nahe, dass ein Großteil der Wechselkursschwankungen im grenzüberschreitenden Handel absorbiert wird. Im Untersuchungszeitraum 1980 bis 2007 reflektieren sich nur 27 Prozent der Wechselkursschocks auf die Schweizer Konsumentenpreise importierter Güter.

KEYWORDS

Import prices, consumer prices, nominal effective exchange rate of the Swiss franc, exchange rate pass-through, terms of trade, Switzerland

Importpreise, Konsumentenpreise, nominaler effektiver Wechselkurs des Schweizer Franken, Überwälzung von Wechselkursveränderungen, Terms of trade, Schweiz

[1] The authors would like to thank an anonymous referee for valuable comments.

1. A Strong Swiss Franc: Pain with no Gain?

The effective exchange rate of the Swiss franc (CHF) reached a historic peak in the summer of 2011. In mid-August 2011 one Euro cost CHF 1.01 compared to CHF 1.62 in August 2011. In September 2011, the Swiss National Bank (SNB) felt impelled to set a minimum threshold at an exchange rate of CHF 1.20 per Euro, in order to mitigate the negative effects of a stronger Swiss franc. This move arose primarily because of the loss of competitiveness of Swiss products on European markets.[2] It partially corrected the massive overvaluation of the Swiss franc. Nevertheless, on January 15th 2015, the Governing Board of the SNB decided that the minimum exchange rate was no longer sustainable and would thus be discontinued. A step that again led to considerable market appreciation of the Swiss franc (Jordan, 2015, p. 2).

However, a strong currency not only is a burden for an economy but also provides an upside – a so-called dividend to the country. One particularly would expect consumers to benefit from potentially declining import prices.[3] But with respect to the appreciation of the Swiss franc this does not appear to be the case. According to the Swiss consumer protection society (SKS – Stiftung für Konsumentenschutz), the prices for imported goods have barely dropped since 2011, as underscored by numerous examples from the retail sector (SKS, 2011). Swissmem – the Association of the Swiss Mechanical and Electrical Engineering Industries – also announced that prices of intermediate products didn't adjust as expected and that currency gains were barely passed through to domestic producers (Hess, 2011). It is still unclear as to where the benefits of the strong Swiss currency are absorbed along the sales chain between foreign producers and domestic consumers. Domestic retailers claim that the problem lies with foreign producers who have been driving up import prices in foreign currencies. A view which to a certain degree is backed by our findings.

Our results show that domestic prices on the import and the consumer level only sluggishly respond to the appreciation of the Swiss franc.[4] The responsiveness of domestic prices arising from exchange rate changes is referred to as exchange rate pass-through (ERPT) (Goldberg and Knetter, 1997). Over the past years, numerous empirical studies on ERPT have been published, among them microanalyses that review the ERPT between firms and locations in various currency zones, and studies that evaluate aggregated data, mostly on a national economy scale (Engel, 1999, McCarthy, 1999, Campa and Goldberg, 2005, Campa et al., 2005).[5]

The paper on hand follows these studies, but first highlights the description of basic statistical data. It is structured as follows: Section 2 distinguishes three levels of ERPT, and Section 3 lists some key empirical findings of the ERPT literature related to Switzerland. Section 4 illustrates the rise of the Swiss franc's effective exchange rate since 1974 and shows that in spite of vigorous appreciation, after the second quarter of 2009, the rising Swiss franc was first accompanied by a decline in terms of trade. The ERPT with respect to import prices in cross-border trade and on the consumer level are considered in Section 5. Section 6 provides a description of the exchange rate responsiveness of overall consumer prices. Section 7 discusses the question of where the exchange rate gains are absorbed in the distribution process between foreign producers and Swiss consumers. The shares of exchange rate shock absorption that occur along the sales chain are explicitly quantified. Section 8 summarizes the key findings.

2. Three Levels of Exchange Rate Pass-Through

Exchange rate pass-through (ERPT) to domestic

[1] The authors would like to thank an anonymous referee for valuable comments.
[2] In May 2012 Jean-Pierre Danthine, the Vice Chairman of the Governing Board of the Swiss National Bank writes, that „[…] the massively overvalued Swiss franc first threatened demand for Swiss exports." (Danthine, 2012).
[3] See for example a statement by Starbatty (2011).
[4] The Swiss Executive Federal Council reacted to this situation by revising the Swiss anti-trust law (EVD, 2011a). The Council justified this competition policy measure explicitly in light of the strong Swiss franc and an inadequate pass-through of exchange rate benefits (EVD, 2011b).The Swiss Competition Commission, which for a long time had been examining potential violations of the antitrust law, welcomed this step (WEKO, 2011). Yet, the Swiss Association of Importers (VSIG, 2011) stated that consumers had already put enough pressure to get sluggish prices moving.
[5] In addition, econometric approaches have been applied, based on multivariate models, cointegration methods, and models that allow mapping of asymmetric and nonlinear relations between currency and pricing variables (Sekine, 2006, Stulz, 2007, De Brandt et al., 2008, Chew et al., 2011, Tressel, 2011).

prices can be viewed from two sides – the import and the export side. In the following, we focus primarily on the import side, where exchange rate changes can potentially impact import prices (IMP), prices of imported consumer goods (CPIIMP) and the overall consumer price index (CPI).[6] Formally, ERPT is measured as the percentage change of domestic prices in local currency resulting from a one percent change in the exchange rate.

IMP-ERPT measures the responsiveness of import prices and, therefore, describes a pass-through phenomenon that occurs in cross-border trade of goods and services. A low IMP-ERPT means that exchange rate changes have little impact on import prices. In the literature, IMP-ERPT is often discussed with respect to issues in competition policy (Taylor, 2000, Stulz, 2007, Campa and Goldberg, 2005). Since summer 2011, Swiss retailers have stated repeatedly that foreign importers of branded goods exploited their market power by absorbing exchange rate gains.[7] This would require that importers have maneuvering room to discriminate prices internationally – a policy that is referred to as pricing-to-market (Dornbusch, 1987).[8]

CPIIMP-ERPT describes the exchange rate pass-through into prices of imported consumer goods traded on the retail level. There are several reasons why CPIIMP-ERPT could be incomplete. One is that imported and domestically produced consumer goods and services cannot be clearly differentiated. Most imported goods are traded, transported, warehoused, financed, and insured within the country. These local components of value added impact the prices of imported goods in the local currency, however, these components are not directly affected by changes in nominal exchange rates. They are higher on the consumer goods level than on the cross-border trade level so that the exchange rate responsiveness shrinks (Burstein et al., 2003). In addition, domestic retailers possibly absorb temporary exchange rate changes in order to minimize their menu costs (Figure 4). This applies in particular with respect to short-run fluctuations. Froot and Klemperer (1989) use U.S. data to demonstrate that permanent nominal exchange rate shocks have a greater impact on local currency prices than temporary shocks. Hedging exchange rate risks, for example, through capital market instruments, can also lower the responsiveness of local currency prices.

CPI-ERPT quantifies the pass-through of nominal exchange rate changes to overall consumer prices, measured as CPI. This ERPT is often discussed in the context of monetary policy issues where a low CPI-ERPT is sometimes referred to as indicating successful monetary policies (Taylor, 2000, Stulz, 2007, Campa and Goldberg, 2005, Gagnon and Ihrig, 2004). Adolfson (2001) stresses that – in reverse causality – the effectiveness of monetary policy is also conditioned upon the degree of CPI-ERPT.[9]

3. Key Empirical Findings of ERPT

ERPT can be measured as short-run, medium-run and long-run coefficients for various trade levels and different categories of goods. In Switzerland-related ERPT literature the following empirical findings and stylized facts are reported: Import prices tend to be more sensitive to changes in the nominal exchange rate than consumer prices. Tressel (2011) looks at the sample period of 1980 to 2010. His long-run IMP-ERPT coefficients for Switzerland range from –0.3 to –0.6, whereas

[6] See Appendix 1.

[7] Auer and Schönle (2012) examine the extent to which market structure can explain incomplete exchange rate pass-through. They evaluate how pass-through rates vary depending on the mass of firms affected by a particular exchange rate shock and the distribution of firms‘ market shares.

[8] If foreign producers set their prices in the importing nation’s currency, their profit or markup floats freely with the exchange rate. The markup rises if the currency of the destination country appreciates. This is in line with the local currency pricing (LCP) model, where nominal exchange rate changes do not evoke price adjustments in the importing country. In other words, no ERPT occurs, at least not over the short run (Engel, 2003). However, according to the producer currency pricing (PCP) model, foreign producers set prices in their own currency, which results in a complete ERPT (Obstfeld, 2002, Obstfeld and Rogoff, 2000b). The question in which currency foreign producers set their prices is crucial in this context (Corsetti, 2007). Fischer et al. (2010) show that in Switzerland big exporters in particular follow pricing to market strategies.

[9] The measurement and interpretation of CPI-ERPT is thus especially complex so that the application of system methods may be more appropriate than single-equation models. Sekine (2006) and Chew et al. (2011) differentiate between a first-stage pass-through and a second-stage pass-through. Under the first stage, they refer to the pass-through of import prices resulting from changes in the nominal exchange rate; the second stage reflects the influence of import prices on the domestic CPI. Amstad and Fischer (2010) analyzed the pass-through of import prices on the Swiss national consumer price index from 1993 to 2008 and determined a coefficient of 0.3.

the corresponding CPI-ERPT coefficients lie between –0.02 and –0.05. Stulz (2007) examines the sample period of 1976 to 2004 and applies a vector autoregressive model to show that after an adjustment period of one year the cumulative IMP-ERPT coefficient is –0.5, while the CPI-ERPT coefficient equals –0.17. The State Secretariat for Economic Affairs (SECO, 2011) by applying the approach of Stulz to the sample period of 1995 to 2011 – derives a pass-through coefficient of –0.4 with respect to import prices.[10] With monthly data that cover the years between 1999 and 2010, Herger (2012) estimates an exchange rate pass-through elasticity onto aggregate import prices of –0.3. McCarthy (1999) and Gagnon and Ihrig (2004) found the CPI-ERPT coefficients for Swiss data to be either minimal or statistically insignificant, whereas the IMP-ERPT coefficients prove slightly higher.

ERPT for import prices differs significantly by category of imported goods and services. A study done by the Deutsche Bundesbank (2008) comes to the result that depending on the category of goods, a one percent change in the nominal exchange rate impacts import prices in local currency by –0.1% up to –0.8%. In particular, imports of raw materials like crude oil and coffee or categories of goods with high shares of raw materials proved more sensitive to exchange rate changes than others. SECO (2011) examines the ERPT for 18 import categories for Switzerland. As expected, the pass-through effects were found to vary with exchange rate changes being relatively well reflected in prices for raw materials (Balastèr, 2011, SECO, 2011).

In the long run, ERPT is generally higher than in the short run. Campa and Goldberg (2005) estimate a short-run IMP-ERPT coefficient of –0.68 and an elasticity of –0.93 after an adjustment period of one year for Switzerland. Adolfson (2001), Heath et al. (2004), de Brandt et al. (2008), and Chew et al. (2011) apply cointegration methods to identify short and long-run effects in various countries. They observe complete IMP-ERPT over the long run. Other analyses focus on the issue of whether the degree of ERPT has changed over time. Taylor (2000), Goldfajn and Werlang (2000) show declining ERPT coefficients. Stulz (2007) concludes that in Switzerland, ERPT dropped during the 1990s where the observed decline was greater for consumer prices than for import prices. For six industrial countries in the sample period of 1974 to 2004, Sekine (2006) also shows a decline in ERPT at both trade levels. He attributes this development to an increasingly effective monetary policy. However, Campa et al. (2005) illustrate that the launching of the common currency in the Euro zone did not cause any structural changes with respect to ERPT.

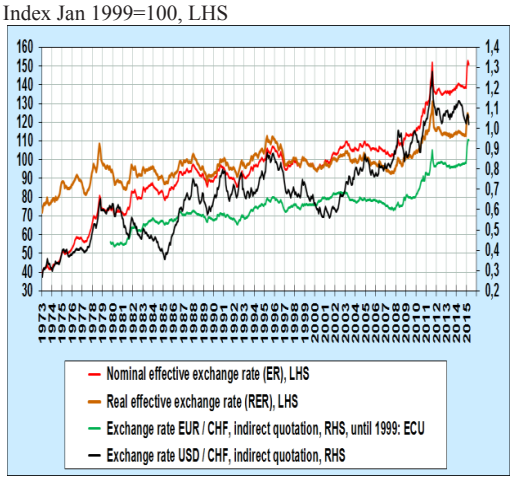
4. Terms of Trade

Since the Bretton Woods system was abandoned in 1973, both the effective exchange rate and key bilateral exchange rates of the Swiss franc have followed a slightly increasing trend (Figure 1, Part a).[11] This trend soared once the financial crisis erupted in the fall of 2007, marked by a rise of the nominal effective exchange rate of more than 49% (October 2007 to August 2011).[12] According to Hildebrand (2011), this appreciation has resulted in a massive overvaluation of the Swiss franc and thus carried the risk of a recession as well as deflationary developments for the Swiss economy.

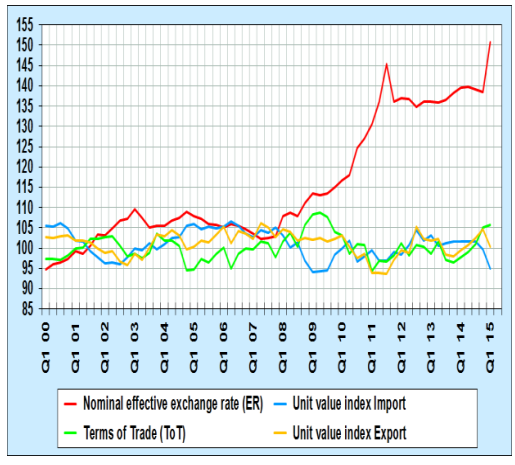
[10] SECO (2011) does not explicitly consider CPIIMP.
[11] See Appendix 1.
[12] This vigorous appreciation was also reflected in the price of the U.S. Dollar and the Euro. The Euro zone countries are by far the major trading partners for Switzerland. The nominal effective exchange rate increased by 36% from October 2007 to December 2014.

Figure 1: Bilateral exchange rates, effective exchange rates of the Swiss franc, terms of trade (goods), export and import prices

(a) Bilateral exchange rates, RHS, effective exchange rate: Index Jan 1999=100, LHS



(b) Effective exchange rate: Index Jan 1999=100, Terms of trade (ToT): previous year=100



Source: Swiss National Bank (SNB), Federal Customs Administration (FCA).

However, since all local consumers of imported goods benefit from potentially declining import prices, the appreciation of the Swiss franc should also show an upside – a so-called dividend to the country. Explicitly, one would expect that the surge of the effective exchange rate increased the amount of import goods the Swiss economy would be able to purchase per unit of export goods. The reason is that a rise in the value of the Swiss franc lowers the domestic prices of imports more than the local currency prices of exports. As a consequence of the appreciation, Swiss terms of trade are supposed to improve. Obstfeld and Rogoff (2000b) attribute this reaction of the terms of trade in the appreciating country to rising relative wages. Yet, under a regime of so-called local currency pricing where responses of import prices to exchange rates changes are assumed to be nil, currency appreciation shows an opposite effect by worsening the country's terms of trade. Finally, in a neoclassical world where all prices adjust flexibly and, thus, the law of one price holds nominal exchange rate changes do not have an effect on the terms of trade.

Figure 1 (Part b) illustrates that the Swiss terms of trade (ToT) did not actually react sharply to the rapid increase of the effective exchange rate. ToT are measured as the relative export prices index (EXP_t) and import prices index (IMP_t) at time t

$$(1) \qquad ToT_t = \frac{1}{ER_t}\frac{EXP_t^*}{IMP_t} = \frac{EXP_t}{IMP_t'}$$

where EXP_t^* denotes export prices in foreign currency and ER_t denotes the nominal effective exchange rate of the Swiss franc (Appendix 1). After the second quarter of 2009, the rising ER_t is first accompanied by a decline in ToT_t that later turns into a stagnation and eventually turns into a slight increase in 2014. The relative stability of Swiss ToT_t could be taken as an indication for the validity of the law of one price (LOP). In this case, the pass-through to import and export prices is expected to be rapid due to an instantaneous adjustment of a small open economy to world prices. But again, a simultaneous decline of EXP_t and IMP_t is not shown in the data (Figure 1, Part b). The stickiness of ToT_t is caused rather by a low responsiveness of the two price indexes, EXP_t and IMP_t.[13]

5. Exchange Rate Pass-Through for Imported Goods and Services

As mentioned above, our focus lies on the import side. In order to learn more about the

[13] See Appendix 1.

responsiveness of the prices of imported goods and services, we look at the relative development of the relevant macroeconomic time series. The data show an unusual response pattern of Swiss import prices after the nominal effective exchange rate shock occurred in the second half of 2007. In our estimation approach, we take this observation into account and examine the question whether the magnitude of pass-through elasticity has changed after 2007. Thus, appropriate procedures for testing parameter stability of the ERPT coefficients are applied.

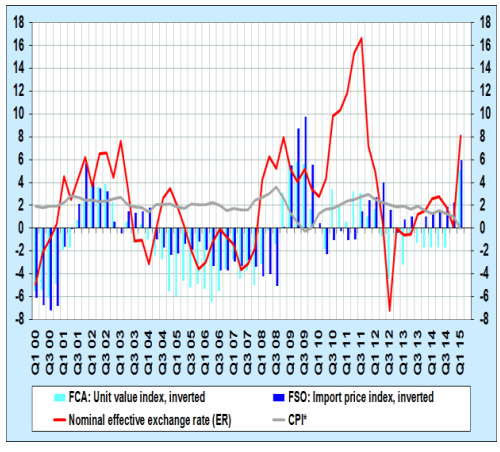
5.1. A first look at Swiss data: Have things recently changed?

Figure 2 (Part a) illustrates the annual rise of the import price index $\Delta \ln IMP_t$ versus $\Delta \ln ER_t$. The red line shows $\Delta \ln ER_t$, which lies above the zero line during appreciation phases and below it during depreciation phases. IMP_t is operationalized in two ways: First, as the import price index of the Federal Statistical Office (FSO) shown as dark columns and secondly, as the unit value index of the Federal Customs Administration (FCA), shown as the lighter column (Appendix 1). Both columns are inverted, which means that a drop in nominal prices points upwards. Figure 2 depicts a certain degree of synchronicity in exchange rate and import price changes. Yet, it is quite evident that the co-movement between $\Delta \ln ER_t$ and $\Delta \ln IMP_t$ has changed in recent years.

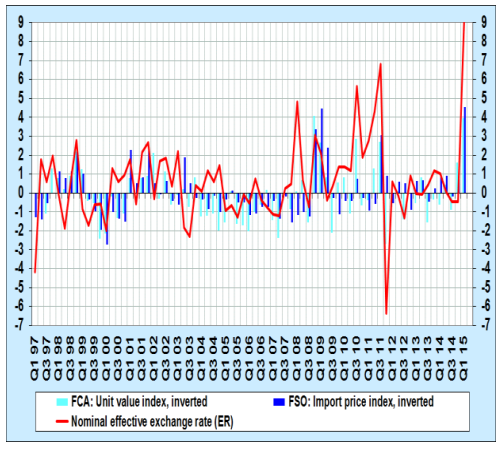
- The depreciation of the Swiss franc in the year 2000 and from 2003 to 2007 was on a comparable order of magnitude as the rise in import prices.
- The Swiss franc appreciated from 2001 to 2003. The annual increase of the nominal effective exchange rate was between 4% and 8%. For eight quarters, a drop of import prices of up to 6% accompanied this development. A second appreciation phase started at the end of 2007, and after a certain time lag the import prices also decreased. In 2009, prices fell by up to 10%, which exceeds the rise in the effective exchange rate – primarily as a result of a major economic downturn (Figure 2, Part a).
- The upward pressure on the Swiss franc rose significantly in 2010 and 2011, leading to year-over-year (y-o-y) rises of the ER_t of up to 16%. At the same time, import prices (based on the import price index of the FSO) increased slightly by 1% y-o-y. The unit value index of the FCA behaved as expected and declined; however, the price reduction remained below 4% (Figure 2, Part a).
- With respect to short-run dynamics, measured as quarter-over-quarter (q-o-q) increase (based on seasonally adjusted data), the counterintuitive movements of import prices are even more evident (Figure 2, Part b). After 2007Q3 ER_t rose – except in 2008Q3 and 2009Q2 – steadily for 14 quarters. Over the same period, the import price index fell for five quarters and rose for 11 quarters.

Figure 2: Import prices, nominal effective exchange rate and foreign consumer price index

(a) y-o-y growth, in percent



b) q-o-q growth, s.a., in percent



Source: Federal Statistical Office (FSO), FCA, SNB, authors' calculations.

5.2. Short- and medium-run IMP-ERPT and CPIIMP-ERPT

According to Campa and Goldberg (2005), the logarithm of the price index of imported goods in foreign currency $\ln IMP_t^*$ can be expressed as the sum of the logarithm of manufacturing costs $\ln MC_t^*$ faced by foreign producers and the markup $\ln MKUP_t^*$ that represents a profit margin. Since $\ln IMP_t = \ln IMP_t^* - \ln ER_t$ holds, the import price index can thus be expressed as

$$(2) \qquad ln\mathrm{IMP}_t = ln\mathrm{MC}_t^* + ln\mathrm{MKUP}_t^* - ln\mathrm{ER}_t$$

Campa and Goldberg (2005) state that the markup $\ln MKUP_t^*$ has a component that is potentially sensitive to changes of the exchange rate. The manufacturing costs $\ln MC_t^*$ depend on prices in the trading partner's markets CPI_t^*. From these considerations Chew et al. (2011) conclude that the price index of imported goods in the local currency

$$(3) \qquad ln\mathrm{IMP}_t = \varepsilon ln\mathrm{ER}_t + \gamma ln\mathrm{CPI}_t^*$$

is a function of the price index of the home country's trading partner's CPI_t^* and ER_t, whereby the coefficient γ reflects the elasticity of import prices with respect to foreign manufacturing costs.[14] The coefficient $\beta^{IMP}=-\varepsilon$ measures the IMP-ERPT, the elasticity of the prices of imported goods with respect to exchange rate variations (Chew et al., 2011). It takes on a value of –1 if the IMP-ERPT is complete.[15] Campa and Goldberg (2005) and Campa et al. (2005) propose a model to estimate short-run and medium-run IMP-ERPT coefficients that can be interpreted as a specification of equation (3):[16]

$$(4) \qquad \Delta ln\mathrm{IMP}_t = \alpha^{\mathrm{IMP}} + \sum_{i=0}^{4} \beta_i^{\mathrm{IMP}} \Delta ln\mathrm{ER}_{t-i} + \sum_{i=0}^{4} \gamma_i^{\mathrm{IMP}} \Delta ln\mathrm{CPI}_{t-i}^* + \varphi^{\mathrm{IMP}} \Delta ln\mathrm{GDP}_t + u_t^{\mathrm{IMP}}.$$

The quarterly change in import prices at time t is explained by the change in the nominal effective exchange rate $\Delta\ln ER_t$, manufacturing costs abroad, and the respective change of these variables in the last four quarters. The increase in foreign manufacturing costs $\Delta\ln MC_t^*$ is operationalized by the increase of the foreign consumer price index $\Delta\ln CPI_t^*$ (depicted by the gray line in Figure 2, Part a). The equation also includes the real gross domestic product $\Delta\ln GDP_t$ as a proxy for short-term demand and medium-term supply conditions in the home market.[17] The short-run pass-through is given by the coefficient $\beta^{IMP,SHORT} = \beta_0^{IMP}$. The medium-run IMP-ERPT elasticity $\beta^{IMP,MEDIUM}$ is given by the sum of the coefficients on the contemporaneous exchange rate and up to four lags of exchange rate terms $\sum_{i=0}^{4} \beta_i^{IMP}$.[18] The short-run and medium-run coefficients are estimated analogously for the price of imported goods at the consumer level ($CPIIMP_t$):

$$(5) \qquad \Delta ln\mathrm{CPIIMP}_t = \alpha^{\mathrm{CPIIMP}} + \sum_{i=0}^{4} \beta_i^{\mathrm{CPIIMP}} \Delta ln\mathrm{ER}_{t-i} + \sum_{i=0}^{4} \gamma_i^{\mathrm{CPIIMP}} \Delta ln\mathrm{CPI}_{t-i}^* + \varphi^{\mathrm{CPIIMP}} \Delta ln\mathrm{GDP}_t + u_t^{\mathrm{CPIIMP}}$$

Appendix 3 presents the estimation results for Switzerland. For the sample period of 1980-2014 the short-run IMP-ERPT coefficient amounts to $\beta^{IMP,SHORT}=-0.20$ whereas the CPIIMP-ERPT coefficient takes on an absolute value slightly lower (–0.16) (column (2) in Table A. and Table B.). Considering the medium-term – that covers an adjustment period of two quarters – the degree of ERPT rises. This is seen in particular for imported goods on the cross-border level, where $\beta^{IMP,MEDIUM}=-0.29$; on

[14] Chew et al. (2011) use the foreign wholesale price index.

[15] Campa and Goldberg (2005) state that $\beta^{IMP}=-1$ supports the hypothesis of producer currency pricing and $\beta^{IMP}=0$ supports the hypothesis of local currency pricing.

[16] Campa and Goldberg (2005, p. 681) do not use the term "medium" but speak of long-run pass-through elasticities.

[17] Campa and Goldberg (2005, p. 681) suggest the real GDP as control variable. The authors argue that "… one should include as the appropriate demand variable an index of income levels across the producer's home market and the destination market for its exports. Because we do not have information on the composition of demand facing exporters in different countries, our proxy here is the GDP of the importing country." For the medium-term, supply conditions which can be expressed as market power or competition intensity can also play a role. Mainly due to lack of available data we assume that GDP is a proxy for these influences, too. Campa and Goldberg (2005, p. 681) conclude that "biased estimates of the pass-through coefficient could arise if foreign wages or GDP are correlated with exchange rates but omitted from the regression". Appendix 1 describes the data used in our estimation.

[18] We formally check the lag length using different information criteria (Akaike info criterion, Schwarz criterion, Hannan-Quinn criterion). For all three equations (4), (5), and (9) only one lag of $\Delta\ln ER$ and $\Delta\ln CPI^*$ is appropriate (Appendix 2).

the consumer level, imported goods and services show a $\beta^{CPIIMP,MEDIUM}$=–0.22.[19] In a second estimation, we consider the subsample period of 1980-2007 that excludes the years following the financial crisis. Over the subsample period, the short-run IMP-ERPT rises slightly to –0.27, whereas the short-run CPIIMP-ERPT increases in absolute value to –0.21. The estimated medium-run coefficients are again lower than the corresponding coefficients estimated over the full sample, with $\beta^{IMP,MEDIUM}$=–0.34; on the consumer level $\beta^{CPIIMP,MEDIUM}$=–0.27.[20]

These estimation results raise the question of whether 2007 indicates a year of significant change with respect to the responsiveness of import prices. The data described in Section 5.1 already suggest that the exchange rate pass through declined after 2007. Following Teräsvirta et al. (2010), Stock and Watson (2007), and Kirchgässner and Wolters (2006), we apply different test procedures to examine whether the hypotheses of parameter constancy can be statistically significantly rejected with respect to the ERPT coefficients. The tests are based on the estimation of equations (4) and (5), specified with one lag. Starting with the single Chow breakpoint test for a break at the date 2007Q3, the null hypothesis of no break is significantly rejected for both equations. In a second step, equations (4) and (5) are estimated with recursive least squares, using ever larger subsets of the sample data. The plot of the recursive residuals signals again some evidence of instability in the parameters of equation (5). The plot shows residuals outside the standard error bands in 2007/2008. However, the Quandt-Andrews breakpoint test – a procedure for a break at an unknown break date – chose 1990Q3 (IMP), respectively 1986Q4 (CPIIMP) as break dates. Multiple breakpoint tests again confirm these two break dates. The observation is backed by the estimation of recursive least squares. The time profile of the recursive $\beta^{IMP,SHORT}$ shows a strong variation in the early 1990s, an indication of instability. Overall, the different tests imply that the IMP-ERPT in Switzerland was considerably higher in the 1980s than in the period of 2000-2014. The re-estimation of equations (4) and (5) for the subsample 1990-2014 provides no significant short-run IMP-ERPT and CPIIMP-ERPT coefficients (Appendix 3, column (7)). In the light of these results the apparently weaker impact of the year 2007 loses relevance.

6. Exchange Rate Pass-Through to CPI

6.1. Volatility and Stationarity of the Real Effective Exchange Rate

A direct comparison of the volatility of Swiss nominal and real effective exchange rates makes it possible to state the responsiveness of the relative prices of all goods and services consumed by Swiss and foreign households. The real effective exchange rate at time t

$$\text{RER}_t = \text{ER}_t \frac{\text{CPI}_t}{\text{CPI}_t^*} \tag{6}$$

is ER_t multiplied by the relative consumer price $\text{CPI}_t/\text{CPI}_t^*$, whereby CPI_t^* denotes the foreign consumer price index. The volatilities of ER_t, RER_t and $\text{CPI}_t/\text{CPI}_t^*$ are measured in terms of the variance of first differences of their log levels. As long as purchasing power parity holds, changes in ER_t are not reflected on the real exchange rate (Obstfeld, 2002). Instead, RER_t is smoothed directly via the adjustment of domestic prices and should thus be markedly less volatile than ER_t. However, in Swiss data, this smoothing effect with respect to the real exchange rate cannot be observed.[21] Rather, the

[19] Our findings for the short-run ERPT are compatible with most empirical literature. Stulz (2007) and Herger (2012) show short-run exchange rate pass-through elasticity onto aggregate Swiss import prices of a magnitude (–0.35 and –0.3 respectively) similar to ours. The CPIIMP-ERPT estimated by Stulz (2007) amounts to –0.27 after three months, which is slightly higher than our result. In contrast, Campa and Goldberg (2005) claim that IMP-ERPT is higher in Switzerland than in most other OECD countries. Their short-run IMP-ERPT coefficient amounts to –0.68, the third highest in a 23 OECD country sample. The sample period in Campa und Goldberg (2005) is 1975-2003. Re-estimating the equation (4) for this time period, our IMP-ERPT turns out to be quite robust with a coefficient of around –0.25. As does Stulz (2007, p. 13), we conclude that the differences in our results are mainly due to different data sets. Campa and Goldberg (2005) use different price series. Their data source for the effective exchange rate indices is International Financial Statistics and the import prices come from the OECD. In our analysis we use Swiss data from the SNB, FSO and SECO (Appendix 1). The fact that Campa and Goldberg (2005) used four lags, whereas our estimation procedure shows – based on different information criteria – an optimal lag length of only one lag, has probably a minor impact. Nevertheless, we estimate our equation (4) with up to four lags of $\Delta\ln\text{ER}_t$ and $\Delta\ln\text{CPI}_t^*$. The short-run IMP-ERPT coefficient varies between –0.20 and –0.22 (Appendix 3, Table A.).

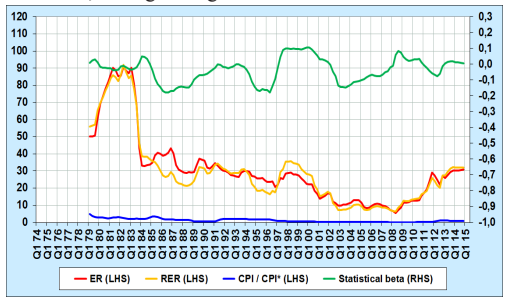
[20] Cointegration tests indicate that IMP-ERPT and CPIIMP-ERPT are asymptotically complete in the sample period of 1980-2007. However, including the post-2007 quarters in the analysis, the long-run CPIIMP-ERPT weakens.

[21] Obstfeld and Rogoff (2000a) speak of a purchasing power parity puzzle.

data indicate that Swiss consumer prices – like import prices – have adjusted only sluggishly to the appreciation of the Swiss franc. In the sample period of 1974-2014, the variance of ER_t turns out to be 37.2, compared with 30.9 for RER_t (Appendix 5).[22] In addition, both variances are computed as rolling values over 20 quarters. Figure 3 shows that the corresponding variance of ER_t over the sample period is only slightly higher than that of RER_t, which indicates a low short-run adjustment of relative consumer prices.

Figure 3: Volatility: Effective exchange rates and relative consumer price index

Variances, rolling averages[1]



[1] Rolling averages of the variance of first-order differences of logarithm values over 20 quarters moving windows.

Source: SNB, Federal Statistical Office (FSO), authors' calculations.

6.2. The responsiveness of relative consumer prices

A simple calculation of $\tilde{\beta}^{CPI}$ coefficients allows us to measure the correlation between nominal effective exchange rates and relative consumer prices. It confirms the low responsiveness of the latter. The variance of RER_t can be decomposed into the variance of ER_t and the variance of the relative price CPI_t/CPI_t^*:

$$\text{(7)} \quad \mathrm{VAR}(\Delta \ln RER_t) = \mathrm{VAR}\left(\Delta \ln \frac{CPI_t}{CPI_t^*}\right) + \mathrm{VAR}(\Delta \ln ER_t)(1+2\tilde{\beta}^{CPI})$$

$$\text{and} \quad \tilde{\beta}^{CPI} = \frac{\mathrm{COV}\left(\Delta \ln \frac{CPI_t}{CPI_t^*}, \Delta \ln ER_t\right)}{\mathrm{VAR}(\Delta \ln ER_t)}$$

The closer $\tilde{\beta}^{CPI}$ approaches a value of –1, the higher is the responsiveness of the relative consumer price index to a change of ER_t. A $\tilde{\beta}^{CPI}$ close to zero points to a weak or highly delayed price adjustment that caused alignment between the volatility of the real and nominal exchange rates. In the sample period of 1974-2014, $\tilde{\beta}^{CPI}$ amounts to –0.15 and the rolling values of $\tilde{\beta}^{CPI}$ range between –0.18 and 0.11 (Figure 3). We get a similar result by focusing on tradable consumer goods only. Equation (8) provides the real effective exchange rate of tradable goods:

$$\text{(8)} \quad RER_t^T = ER_t \left(\frac{EUR}{CHF}\right) \frac{CPI_t^T}{CPI_t^{T,G}}.$$

It equals the nominal effective exchange rate of the Swiss franc versus the Euro, ER_t (EUR/CHF), multiplied by the relative price index for tradable goods of Switzerland CPI_t^T and Germany $CPI_t^{T,G}$.[23] We expect the prices of tradable goods to be more sensitive to changes in ER_t than the prices of non-tradable goods. However, focusing on tradable goods does not raise the ERPT coefficient. Compared to the variance of ER_t, the variance of RER_t^T proves even higher (Appendix 5). For the subsample period of 2006 to 2014, $\tilde{\beta}^{CPI,T}$ for tradable goods turns out to be –0.01 and thus shows a slightly lower magnitude than $\tilde{\beta}^{CPI}$ (–0.02). Even though the prices of non-tradable goods should be less affected by changes in ER_t, their exclusion does not raise the coefficients.

A further possible explanation for the low value of $\tilde{\beta}^{CPI}$ could be that domestic prices adjust only with a delay. This means that after a nominal exchange rate shock a considerable time lag follows until a marked pass-through occurs. Due to the stickiness of domestic prices, real exchange rates follow the movement of ER_t in the short run and only gradually converge back to their original level.[24] In the long run, however, adjustment should be complete, indicated by a long-run coefficient $\beta^{CPI,LONG}=-1$. This condition holds if the time series RER_t is asymptotically stationary, which can be verified through the application of a simple unit root test (Kim, 1990, Glen, 1992, Rogoff, 1996). In the sample period of 1973 to 2014,

[22] If the indices follow different trends, applying the variance as a measure of volatility could lead to false conclusions. In this case, the mean squared error $MSE_t^i=(\sigma_t^i)^2+(\mu_t^i)^2$ serves as a superior volatility measure (Engel, 1999). It represents the sum of the variance $(\sigma_t^i)^2$ and the square of the arithmetic mean μ_t^i of the respective index. Since ER_t follows a steeper trend, the volatility discrepancy between MSE_t^{ER} (44.8) and MSE_t^{RER} (31.6) is boosted in the sample period. The mean squared errors computed as rolling values over 20 quarters are roughly the same for both indices.

[23] See Appendix 1.

[24] Since CPI_t^* represents the price development of 24 Swiss trading partners it is not expected to respond sensitively to changes of the price of the Swiss franc.

the hypothesis of a unit root in the time series of RER_t can be rejected, which is also valid for the subsample period of 1973 to 2007 (Appendix 4). This supports the hypothesis that CPI-ERPT becomes complete after a period of sufficient length.

6.3. The responsiveness of CPI

The following estimation results give further evidence that the stickiness of CPI_t/CPI_t^* indicates a low CPI-ERPT. To estimate short-run and medium-run CPI-ERPT coefficients we once again apply the model suggested by Goldberg and Campa (2005) and Campa et al. (2005):[25]

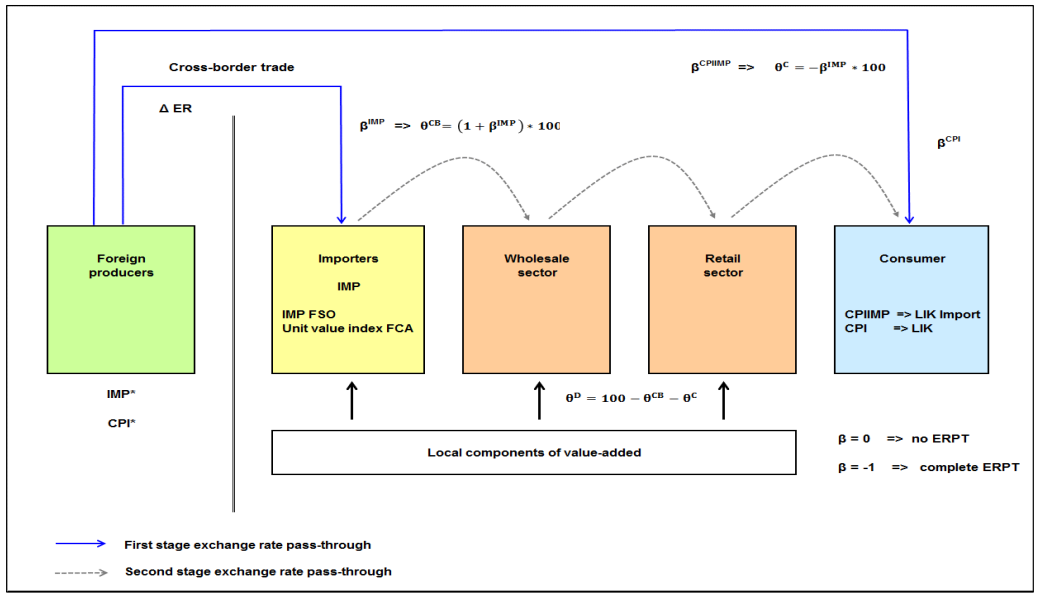
$$\Delta \ln CPI_t = \alpha^{CPI} + \sum_{i=0}^{4} \beta_i^{CPI} \Delta \ln ER_{t-i} + \sum_{i=0}^{4} \gamma_i^{CPI} \Delta \ln CPI^*_{t-i} + \varphi^{CPI} \Delta \ln GDP_t + u_t^{CPI} \quad (9)$$

The estimation results – presented in Appendix 3 – show that over the short run the Swiss CPI does not respond much to exchange rate changes. The CPI-ERPT coefficient for both sample periods is small. The coefficient $\beta^{CPI,SHORT}$ amounts to –0.05 in the period of 1980-2014 and to –0.06 in the period of 1980-2007. Considering the medium run – that again covers an adjustment period of two quarters – the degree of ERPT rises only slightly. In the full sample period, CPI-ERPT shows a coefficient $\beta^{CPI,MEDIUM}$= –0.09, in the subsample the coefficient adjusts to –0.08. According to Stulz (2007), the short-run CPI-ERPT coefficient amounts to –0.09 after three months. His CPI-ERPT coefficient after 6 months is –0.14.[26]

7. Who Benefits from the Strong Swiss Franc?

Our estimation results show that over the short and medium run, even the prices for imported goods exhibit only a minor ERPT. Exchange rate shocks are thus partially absorbed in the distribution process of imported goods and services and do not reach the Swiss consumer. Shock absorption can potentially occur in cross-border trade as well as in domestic trade, which includes the wholesale and the retail sector (Figure 4).

Figure 4: Levels of Exchange Rate Pass-Through and Shock Absorption



To determine where exchange rate shocks are actually absorbed along the sales chain, we use simple variance decomposition starting with the identity

[25] We formally check the lag length using different information criteria. For equation (9) only one lag of ΔlnER and $\Delta \ln CPI^*$ is appropriate (Appendix 2).

[26] Engle and Granger (1987) tests are run for consumer prices. In both sample periods, no cointegration relationship between CPI and ER is found. Equation (9) is further re-estimated in error-correction form. No significant cointegration relationship is found either. Both test results claim that even long-run ERPT for consumer prices is not complete. This finding is in line with earlier results from McCarthy (1999), Gagnon and Ihrig (2004), and Tressel (2011).

(10) $$ER_t = \frac{ER_t}{IMP_t} \frac{IMP_t}{CPIIMP_t} CPIIMP_t$$

From equation (10) we take logs and first differences, multiply both sides by $\Delta lnER_t$ and introduce expectation values to obtain

(11)
$$E(\Delta lnER_t^2) = E(\Delta lnER_t^2)+E(\Delta lnIMP_t\,\Delta lnER_t) - E((\Delta lnIMP_t - \Delta lnCPIIMP_t)\Delta lnER_t) - E(\Delta lnCPIIMP_t\Delta lnER_t)$$

This expression is rewritten in terms of variances (VAR) and covariances (COV):

(12)
$$1=1+\underbrace{\frac{COV(\Delta lnIMP_t,\Delta lnER_t)}{VAR(\Delta lnER_t)}}_{1+\tilde{\beta}_t^{IMP}} - \frac{COV((\Delta lnIMP_t-\Delta lnCPIIMP_t),\Delta lnER_t)}{VAR(\Delta lnER_t)} - \underbrace{\frac{COV(\Delta lnCPIIMP_t,\Delta lnER_t)}{VAR(\Delta lnER_t)}}_{\tilde{\beta}_t^{CPIIMP}}$$

where the coefficients $\tilde{\beta}_t^{IMP}$ and $\tilde{\beta}_t^{CPIIMP}$ represent simple OLS estimators (Asdrubali et al., 1996).[27]

These estimators can be used to derive the percentage of exchange rate shocks absorbed on different levels and thus allow the quantification of exchange rate shock absorption in cross-border and domestic trade. Not absorbed in the distribution process and thus reflected in consumer prices is $\theta_t^C=(-\tilde{\beta}_t^{CPIIMP})*100$.

The fraction of the shock that is absorbed in cross-border trade and not reflected in domestic import prices is given by $\theta_t^{CB}=(1+\tilde{\beta}_t^{IMP})*100$. Since IMP accounts for purchasing prices paid by importers, the coefficient θ_t^{CB} measures the share of exchange rate changes absorbed by foreign producers or traders. The percentage share absorbed on domestic trade levels is given by the residual $\theta_t^D=100-\theta_t^{CB}-\theta_t^C$. How this percentage share is distributed among the domestic importers, wholesaler and retailers of imported goods cannot be determined due to a lag of according price statistics.

Table 1: Percentage of the exchange rate shock absorbed by different levels of merchandise distribution

(a) short run (one quarter)

Level of merchandise distribution		**1980:Q1 – 2014:Q4**		**1980:Q1 – 2007:Q3**	
		absorption	**reflection**	**absorption**	**reflection**
Cross-border trade level	$\theta^{CB} = (1+ \beta^{IMP,SHORT})*100$	80.2%		73.2%	
Domestic trade level	$\theta^{D} = 100 - \theta^{CB} - \theta^{C}$	4.3%		6.3%	
Consumers	$\theta^{C} = -\beta^{CPIIMP,SHORT} *100$		**15.5%**		**20.5%**
Sum of absorption and reflection		100%		100%	

(b) medium run (two quarters)

Level of merchandise distribution		**1980:Q1 – 2014:Q4**		**1980:Q1 – 2007:Q3**	
		absorption	**reflection**	**absorption**	**reflection**
Cross-border trade level	$\theta^{CB} = (1+ \beta^{IMP,MEDIUM})*100$	70.6%		65.9%	
Domestic trade level	$\theta^{D} = 100 - \theta^{CB} - \theta^{C}$	7.0%		7.1%	
Consumers	$\theta^{C} = -\beta^{CPIIMP,MEDIUM} *100$		**22.4%**		**27.0%**
Sum of absorption and reflection		100%		100%	

The specific fractions of shock absorption depend on the responsiveness of IMP and CPIIMP. The results depicted in Table 1 are therefore derived using ERPT coefficients instead of statistical OLS estimators. Once again, the two sample periods of 1980 to 2007 and 1980 to 2014 are differentiated. For both periods, short-run and medium-run absorption is derived based on the corresponding ERPT coefficients. Before 2007, 21% of the exchange rate shocks were reflected in the domestic CPI for imported goods over the short run. This means that 79% of the exchange rate shocks did not reach the consumer level. Almost 73% were absorbed in cross-border trade and 6% by the domestic trade sector (Table 1, Part a). In the medium run, the CPIIMP-ERPT amounts to 27%. The remaining 73% of the exchange rate shock were absorbed. 66% went to foreign producers or traders and 7% were absorbed by the domestic trade level.

[27] See Campa et al. (2005).

Considering the full sample period of 1980-2014, the shock absorption in the distribution process shows generally higher. In the short (medium) run only 16% (22%) of shocks were reflected in consumer prices for imported goods. The percentage amount absorbed in domestic trade fell slightly to 4% (7%). Simultaneously, the proportion absorbed in cross-border trade rose to 80% over the short run and to nearly 71% over the medium run.

8. Conclusions

The effective exchange rate of the Swiss franc has soared since the outbreak of the financial crisis in 2007. As price statistics show, the appreciation led only to a minor adjustment of the Swiss terms of trade, attributed to sticky import prices as well as export prices measured in Swiss francs respectively. Focusing on the import side, the low responsiveness of the Swiss terms of trade indicates a weak ERPT with respect to Swiss import prices. To quantify pass-through effects we apply an approach proposed by Campa and Goldberg (2005), which allows us to distinguish short-run and medium-run effects. We further look at different price levels: import prices (IMP-ERPT), consumer prices for imported goods (CPIIMP-ERPT), and the overall consumer price index (CPI-ERPT). Our results can be summarized as follows:

ERPT is higher with respect to import prices than with respect to consumer prices. In the period of 1980-2007, the short-run IMP-ERPT coefficient – measuring the adjustment within a quarter – amounts to –0.27, whereas the pass-through into the prices of imported goods on the consumer level is slightly lower. The short-run CPI-ERPT coefficient is –0.06. ERPT increases when the adjustment period is extended from one quarter to 6 months. Our analyses further suggests a moderate slowdown of the pass-through after 2007, marked by declining ERPT coefficients in the prolonged sample period of 1980-2014. However, different tests of parameter stability applied signal no strong evidence for a structural break in 2007.

Our findings allow inferences of where exchange rate gains are absorbed along the sales chain between foreign producers and Swiss consumers. From 1980 to 2007, the largest share of exchange rate shocks is absorbed in cross-border trade, leaving only about 7% to domestic traders. Swiss consumers of imported goods felt around 27% of initial exchange rate shocks. In the prolonged sample period of 1980 to 2014, foreign producers and traders absorbed even more of the exchange rate shock. These findings underscore the position taken by Swiss retailers, whereby the benefits of the currency appreciation get caught at the Swiss border.

9. References

Adolfson, Malin (2001): Monetary Policy with Incomplete Exchange Rate Pass-Through. Sveriges Riksbank, Working Paper Series No. 127 (September 2001).

Amstad, Marlene; Fischer, Andreas M. (2010): "Monthly Pass-Through Ratios". In: *Journal of Economic Dynamics and Control* 34 (7), pp. 1202–1213.

Asdrubali, Pierfederico; Sørensen, Bent. E.; Yosha, Oved (1996): "Channels of Interstate Risk Sharing: United States 1963–1990". In: *Quarterly Journal of Economics* 111 (4), pp. 1081–1110.

Auer, Ralph A.; Schoenle, Raphael S. (2012): Market Structure and Exchange Rate Pass-Through. Federal Reserve Bank of Dallas, Globalization and Monetary Policy Institute, Working Paper No. 130 (February 2012).

Balastèr, Peter (2011): „Weitergabe von Einkaufsvorteilen aufgrund der Frankenstärke". In: *Die Volkswirtschaft (Das Magazin für Wirtschaftspolitik),* Eidgenössisches Volkswirtschaftsdepartement und Staatssekretariat für Wirtschaft (Ed.) (No. 11), pp. 56–60.

Burstein, Ariel T.; Neves, João C.; Rebelo, Sergio (2003): "Distribution Costs and Real Exchange Rate Dynamics During Exchange-rate Based Stabilizations". In: *Journal of Monetary Economics* 50 (6), pp. 1189–1214.

Bundesamt für Statistik (BFS)/Swiss Federal Statistical Office (2011): Produzenten- und Importpreisindex (Dezember 2010=100). Methodenübersicht. Neuchâtel (February 2011).

Campa, José M.; Goldberg, Linda S. (2005): "Exchange Rate Pass-Through into Import Prices". In: *The Review of Economics and Statistics* 87 (4), pp. 679–690.

Campa, José M.; Goldberg, Linda S.; González-Mínguez, José M. (2005): Exchange-Rate Pass-Through to Import Prices in the Euro Area. The National Bureau of Economic Research (NBER) Working Paper No. 11632 (September 2005).

Chew, Joey; Tan, Siang M.; Ouliaris, Sam (2011): Exchange Rate Pass-Through Over the Business Cycle in Singapore. International Monetary Fund (IMF) Working Paper No. 11/141 (June 2011).

Corsetti, Giancarlo (2007): New Open Economy Macroeconomics. Centre for Economic Policy Research (CEPR) Discussion Paper No. 6578. London (November 2007).

Danthine, Jean-Pierre (2012): Monetary Policy is not Almighty. Speech by Jean-Pierre Danthine, Vice Chairman of the Governing Board of the Swiss National Bank, at the Journée Solutions Bancaires, Geneva, May 31, 2012.

De Brandt, Olivier; Banerjee, Anindya; Koźluk, Tomasz (2008): "Measuring Long-Run Exchange Rate Pass-Through". In: *Economics (The Open-Access, Open-Assessment E-Journal);* Special Issue „Recent Developments in International Money and Finance" 2 (2008-6).

Deutsche Bundesbank (2008): Gesamtwirtschaftliche Effekte realer Wechselkursänderungen. In: Monatsberichte der Deutschen Bundesbank (März 2008). Frankfurt am Main, Germany, pp. 35–61.

Dornbusch, Ruediger (1987): "Exchange Rates and Prices". In: *American Economic Review* 77 (1), pp. 93–106.

Engel, Charles (1999): "Accounting for U.S. Real Exchange Rate Changes". In: *Journal of Political Economy* 107 (3), pp. 507–537.

Engel, Charles (2002): "Expenditure Switching and Exchange Rate Policy". The National Bureau of Economic Research (NBER) Macroeconomics Annual 2002 (Vol. 17), pp. 231–272.

Engle, Robert F.; Granger, Clive W.J. (1987): "Co-Integration and Error Correction: Representation, Estimation, and Testing". In: *Econometrica* 55 (2), pp. 251–276.

EVD - Eidgenössisches Volkswirtschaftsdepartement (2011a): Teilkartellverbot mit Rechtfertigungsmöglichkeit: Anpassung von Artikel 5 Kartellgesetz gemäß Entscheid des Bundesrates vom 17. August 2011. Bern.

EVD - Eidgenössisches Volkswirtschaftsdepartement (2011b): Konferenzielle Vernehmlassung zur Revision des Kartellgesetzes. Bern (September 23rd, 2011).

FCA- Swiss Federal Customs Administration/ Eidgenössische Zollverwaltung (EZV); Sektion Diffusion und Analysen (2006): Schweizerische Aussenhandelsindizes. Benutzerleitfaden. Bern (May 2006).

Fischer, Andreas M.; Lutz, Matthias; Wälti, Manuel (2010): "Pricing-To-Markets and Firm Size: Survey Evidence from Swiss Exporters". In: *Aussenwirtschaft* 65 (March 1), pp. 73–88.

Froot, Kenneth P.; Klemperer, Paul D. (1989): "Exchange Rate Pass-Through when Market Share Matters". In: *The American Economic Review* 79 (4), pp. 637–654.

FSO - Swiss Federal Statistical Office/Bundesamt für Statistik (BFS) (2011): Produzenten- und Importpreisindex (Dezember 2010=100). Methodenübersicht. Neuchâtel (February 2011).

Gagnon, Joseph E.; Ihrig, Jane (2004): Monetary Policy and Exchange Rate Pass Through. Board of Governors of the Federal Reserve System, International Finance Discussion Papers Number 704. Revised Version.

Glen, Jack D. (1992): "Real Exchange Rates in the Short, Medium, and Long Run". In: *Journal of International Economics* 33 (1-2), pp. 147–166.

Goldberg, Pinelopi K.; Knetter, Michael M. (1997): "Goods Prices and Exchange Rates: What Have We Learned?" In: *Journal of Economic Literature* 35 (3), pp. 1243–1272.

Goldfajn, Ilan; R. da C. Werlang, Sergio (2000): The Pass-through from Depreciation to Inflation: A Panel Study. Department of Economics, Pontificia Universidade Catolica (PUC), Rio de Janeiro, Brazil. Working Paper No. 423.

Heath, Alexandra; Roberts, Ivan; Bulman, Tim (2004): "Inflation in Australia: Measurement and Modelling". In: Kent, Christopher; Guttmann Simon (Ed.): *The Future of Inflation Targeting – Conference 2004,* Economic Group, Reserve Bank of Australia, Sidney, pp. 167-207.

Herger, Nils (2012): "Exchange Rates and Import Price in Switzerland". In: *Swiss Journal of Economics and Statistics* 148 (III), pp. 381–407.

Hildebrand, Philipp (2011): Short Statement with Regard to the Introduction of a Minimum Swiss Franc Exchange Rate against the Euro. Philipp Hildebrand, Chairman of the Governing Board of the Swiss National Bank, Zurich, Switzerland, September 9, 2011.

Jordan, Thomas J. (2015): SNB Monetary Policy after the Discontinuation of the Minimum Exchange Rate. Chairman of the Governing Board of the Swiss National Bank. 107th Ordinary General Meeting of Shareholders of the Swiss National Bank, Bern, April 24, 2015.

Kellermann, Kersten; Schlag, Carsten-Henning (2011): Frankenstärke und Importpreisreagibilität: Kurz-, mittel- und langfristige Effekte. In: Konjunkturforschungsstelle Liechtenstein (KOFL) (Ed.): KOFL Working Papers. Vaduz, Liechtenstein (October 2011, No. 10).

Kim, Yoonbai (1990): "Purchasing Power Parity in the Long Run. A Cointegration Approach". In: *Journal of Money, Credit and Banking* 22 (4), pp. 491–503.

Kirchgässner, Gebhard; Wolters, Jürgen (2006): Einführung in die moderne Zeitreihenanalyse. München: Verlag Franz Vahlen (WiSo-Kurzlehrbücher).

MacKinnon, James G. (1996): "Numerical Distribution Functions for Unit Root and Cointegration Tests". In: *Journal of Applied Econometrics* 11 (6), pp. 601–618.

McCarthy, Jonathan (1999): Pass-Through of Exchange Rates and Import Prices to Domestic Inflation in Some Industrialized Economies. BIS Working Papers No. 79. Basel, Switzerland.

Obstfeld, Maurice (2002): Exchange Rates and Adjustment: Perspectives from the New Open-Economy Macroeconomics. Keynote Speech. Bank of Japan. In: Monetary and Economic Studies. Special Edition (December 2002), pp. 23–46.

Obstfeld, Maurice; Rogoff, Kenneth S. (2000a): "The Six Major Puzzles in International Macroeconomics: Is there a Common Cause?" In: Ben Bernanke und Kenneth

S. Rogoff (Ed.): National Bureau of Economic Research (NBER) Macroeconomics Annual 2000. Cambridge, Mass., USA: MIT Press, pp. 339–390.

Obstfeld, Maurice; Rogoff, Kenneth S. (2000b): "New Directions for Stochastic Open Economy Models". In: *Journal of International Economics* 50, pp. 117–153.

Rogoff, Kenneth S. (1996): "The Purchasing Power Parity Puzzle". In: *Journal of Economic Literature* 34 (2), pp. 647–668.

SECO - State Secretariat for Economic Affairs [Staatssekretariat für Wirtschaft] (2011): Weitergabe von Einkaufsvorteilen aufgrund der Frankenstärke („Transparenzstudie"). Working Paper. Bern (November 2011).

SKS - Stiftung für Konsumentenschutz (2011): Starker Franken: SKS fordert einen runden Tisch, Preise von Importgütern müssen endlich deutlich sinken. Medienmitteilung vom 14. Juli 2011. Bern, Switzerland.

Starbatty, Joachim (2011): „Geldpolitik ist kein starkes Schwert in Krisenzeiten". In: Neue Zürcher Zeitung (NZZ), August 10, 2011.

Stock, James H.; Watson, Mark W. (2007): Introduction to Econometrics. 2nd Edition. Boston: Pearson/Addison Wesley International Edition (The Addison-Wesley series in economics).

Stulz, Jonas (2007): Exchange Rate Pass-Through in Switzerland: Evidence from Vector Autoregressions. Swiss National Bank Economic Studies No. 4.

Taylor, John (2000): "Low Inflation, Pass-Through, and the Pricing Power of Firms". In: *European Economic Review* 44 (7), pp. 1389–1408.

Teräsvirta, Timo; Tjøstheim, Dag; Granger, Clive W. J (2010): Modelling Nonlinear Economic Time Series. Oxford, New York: Oxford University Press (Advanced texts in econometrics).

Tressel, Thierry (2011): Impact of Exchange Rate Movements on Exports Performance and Consumer Prices. In: International Monetary Fund (IMF) (Ed.): IMF Country Report No. 11/116. Switzerland: Selected Issues Papers, Washington, D.C. (May 2011).

VSIG Handel Schweiz (2011): Die Chance der Krise. Basel, Switzerland (August 9, 2011).

WEKO - Eidgenössische Wettbewerbskommission (2011): Unvollständige Weitergabe von Wechselkursvorteilen. Bern, Switzerland (August 2, 2011).

Appendix

Appendix 1: Data

CPI = Consumer price index

Swiss consumer price index ("Landesindex der Konsumentenpreise – LIK"), December 2010=100, seasonally adjusted. Source: Federal Statistical Office (FSO), KOF Swiss Economic Institute.

CPI* = Foreign consumer price index, trade weighted

The foreign trade weighted consumer price index is calculated as CPI*=CPI(ER/RER).

CPI/CPI* = Relative consumer price index

The relative consumer price index (CPI/CPI*) is calculated as RER/ER. Source: Swiss National Bank (SNB).

CPIIMP = Consumer price index, sub-index imported goods

Swiss consumer price index, sub-index imported goods, December 2010=100, seasonally adjusted. Source: FSO, KOF Swiss Economic Institute.

ER = Nominal effective exchange rate of the Swiss franc

Nominal effective exchange rate of the Swiss franc versus 24 trading partners, not seasonally adjusted. Source: SNB, KOF Swiss Economic Institute.

GDP = Real gross domestic product

Real gross domestic product, at prices of the preceding year, chained values ("annual overlap"), reference year 2000. Source: State Secretariat for Economic Affairs (SECO).

IMP = Import price index

Import price index (goods and services) is published by the Federal Statistical Office (FSO). For more details see FSO (2011, p. 12), December 2010=100. Source: FSO, KOF Swiss Economic Institute.

Unit value index

Unit value index (goods) is published by the Swiss Federal Customs Ad-mination (FCA). For more details see FCA (2006), Previous year=100. Source: FCA.

RER = Real effective exchange rate of the Swiss franc

The real effective exchange rate of the Swiss franc versus 24 trading partners is calculated by the Swiss National Bank (SNB). For deflating, the SNB uses the Swiss consumer price index ("Landesindex der Konsumentenpreise", LIK) und the consumer price indices from the trading partners, Source: SNB, KOF Swiss Economic Institute.

RER^T = Real effective exchange rate of the Swiss franc against Germany (tradable goods)

The nominal exchange rate of the Swiss franc versus the Euro deflated by the relative prices of Swiss and German tradable goods. The latter is not available. For calculation the sub-indices "Index ohne Wohnungsmiete", that covers 81.1% of the LIK and the German "Gesamtindex ohne Nettokaltmieten" are used (Schlag and Kellermann, 2011). It covers 76.3% of the German CPI, Source: FSO, Federal Statistical Office of Germany.

ToT = Terms of Trade

The Swiss terms of trade index is published by the FCA. It is calculated as the price of exports in relation to the price of imports, based on unit value indices of the FCA, Source: FCA.

Appendix 2: Analysis of the lag structure to estimate short-run and medium-run ERPT coefficients

Sample period: 1980:Q1 to 2014:Q4

Lags (quarters)	Information criteria for …								
	estimation of eq. (4)			estimation of eq. (5)			estimation of eq. (9)		
	AIC	SC	HQ	AIC	SC	HQ	AIC	SC	HQ
0	-6.0956	-6.0111	-6.0613	-6.5134	-6.4389	-6.4891	-8.2095	-8.1250	-8.1751
1	-6.1035	-5.9768	-6.0519	-6.5218	-6.3952	-6.4704	-8.2312	-8.1045	-8.1797
2	-6.0811	-5.9122	-6.0125	-6.4988	-6.3299	-6.4301	-8.2171	-8.0482	-8.1484
3	-6.0888	-5.8777	-6.0030	-6.4906	-6.2795	-6.4048	-8.1913	-7.9801	-8.1055
4	-6.0719	-5.8187	-5.9690	-6.4769	-6.2236	-6.3739	-8.1625	-7.9092	-8.0596
5	-6.0463	-5.7507	-5.9262	-6.4709	-6.1753	-6.3508	-8.1561	-7.8605	-8.0359
6	-6.0271	-5.6893	-5.8898	-6.4602	-6.1225	-6.3229	-8.1696	-7.8318	-8.0324
7	-6.0007	-5.6207	-5.8463	-6.4462	-6.0662	-6.2918	-8.1734	-7.7934	-8.0189
8	-5.9981	-5.5758	-5.8265	-6.4261	-6.0039	-6.2545	-8.1831	-7.7609	-8.0116
9	-5.9742	-5.5097	-5.7855	-6.4111	-5.9466	-6.2223	-8.2002	-7.7358	-8.0115
10	-5.9466	-5.4399	-5.7407	-6.4065	-5.8999	-6.2007	-8.2281	-7.7161	-8.0169
11	-5.9310	-5.3821	-5.7079	-6.3850	-5.8361	-6.1619	-8.2172	-7.6683	-7.9941
12	-5.9101	-5.3189	-5.6698	-6.3709	-5.7798	-6.1307	-8.1991	-7.6079	-7.9589

Information criteria: Akaike info criterion (AIC), Schwarz criterion (SC), Hannan-Quinn criterion (HQ).

Appendix 3: Estimation results of the short, medium and long-run IMP-ERPT, CPIIMP-ERPT and CPI-ERPT

A. IMP-ERPT Eq. (4): dependent variable: $\Delta lnIMP_t$ t-values in parenthesis

variable	Campa and Goldberg (2005) specification (different lags)						
	1980-2014					1980-2007	1990-2014
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
constant	-0.008 (-4.7)	-0.007 (-4.2)	-0.006 (-3.6)	-0.005 (-3.0)	-0.006 (-2.7)	-0.007 (-2.9)	-0.009 (-4.3)
$\Delta lnER_t$	-0.220 (-5.0)	-0.198 (-4.4)	-0.202 (-4.5)	-0.214 (-4.7)	-0.212 (-4.7)	-0.268 (-5.0)	-0.033 (-0.7)
$\Delta lnER_{t-1}$		-0.096 (-2.2)	-0.085 (-1.9)	-0.088 (-1.9)	-0.957 (-2.1)	-0.073 (-1.4)	-0.096 (-2.1)
$\Delta lnER_{t-2}$			-0.028 (-0.6)	-0.011 (-0.3)	-0.010 (-0.2)		
$\Delta lnER_{t-3}$				-0.039 (-0.8)	-0.033 (-0.7)		
$\Delta lnER_{t-4}$					0.006 (0.1)		
$\Delta lnCPI^*_t$	0.709 (4.3)	0.824 (3.3)	0.899 (-3.4)	1.047 (3.7)	1.063 (3.7)	0.665 (2.4)	0.742 (2.9)
$\Delta lnCPI^*_{t-1}$		-0.118 (-0.6)	-0.039 (-0.2)	0.118 (0.4)	0.227 (0.7)	-0.041 (-0.2)	-0.039 (-0.2)
$\Delta lnCPI^*_{t-2}$			-0.199 (-0.8)	0.025 (0.1)	0.077 (0.3)		
$\Delta lnCPI^*_{t-3}$			-0.557 (-2.1)	-0.422 (-1.4)			
$\Delta lnCPI^*_{t-4}$				-0.331 (-1.2)			
$\Delta lnGDP_t$	0.787 (5.4)	0.742 (5.1)	0.692 (4.4)	0.600 (3.6)	0.578 (3.4)	0.576 (3.5)	1.031 (6.7)
Adj. R^2	0.326	0.359	0.354	0.368	0.365	0.319	0.405
$\beta^{IMP,SHORT}$	-0.220	-0.198	-0.202	-0.214	-0.212	-0.268	0
$\beta^{IMP,MEDIUM}$		-0.294	-0.315	-0.352	-0.350	-0.341	0

B. CPIIMP-ERPT Eq. (5): dependent variable: $\Delta lnCPIIMP_t$ t-values in parenthesis

variable	Campa and Goldberg (2005) specification (different lags)						
	1980-2014					1980-2007	1990-2014
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
constant	-0.005 (-3.5)	-0.004 (3.0)	-0.004 (-2.6)	-0.003 (-2.1)	-0.003 (-1.9)	-0.002 (-1.5)	-0.001 (-4.4)
$\Delta lnER_t$	-0.171 (-4.8)	-0.155 (-4.2)	-0.157 (-4.3)	-0.165 (-4.4)	-0.163 (-4.6)	-0.205 (-5.0)	-0.042 (-1.1)
$\Delta lnER_{t-1}$		-0.066 (-1.8)	-0.058 (-1.6)	-0.059 (-1.6)	-0.067 (-1.8)	-0.065 (-1.6)	-0.037 (-1.0)
$\Delta lnER_{t-2}$			-0.014 (-0.4)	-0.003 (-0.1)	-0.001 (-0.1)		
$\Delta lnER_{t-3}$				-0.019 (-0.5)	-0.016 (-0.4)		
$\Delta lnER_{t-4}$					0.016 (0.4)		
$\Delta lnCPI^*_t$	0.906 (6.8)	1.001 (4.9)	1.064 (4.9)	1.163 (4.9)	1.184 (5.1)	0.722 (3.5)	1.203 (5.6)
$\Delta lnCPI^*_{t-1}$		-0.118 (-0.6)	-0.025 (-0.1)	0.068 (0.3)	0.176 (0.7)	0.083 (0.4)	0.035 (0.2)
$\Delta lnCPI^*_{t-2}$			-0.177 (-0.8)	-0.035 (-0.2)	0.001 (0.1)		
$\Delta lnCPI^*_{t-3}$				-0.350 (-1.6)	-0.237 (-1.0		
$\Delta lnCPI^*_{t-4}$					-0.292 (-1.3)		
$\Delta lnGDP_t$	0.186 (1.6)	0.154 (1.3)	0.119 (0.9)	0.068 (0.5)	0.051 (0.4)	-0.064 (-0.5)	0.447 (3.5)
Adj. R^2	0.326	0.335					
$\beta^{IMP,SHORT}$	**-0.171**	**-0.155**	**-0.157**	**-0.165**	**-0.163**	**-0.205**	**0**
$\beta^{IMP,MEDIUM}$		**-0.221**	**-0.228**	**-0.246**	**-0.231**	**-0.270**	**0**

C. CPI-ERPT Eq. (9): dependent variable: $\Delta lnCPI_t$ t-values in parenthesis

variable	Campa and Goldberg (2005) specification (different lags)						
	1980-2014					1980-2007	1990-2014
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
constant	0.002 (0.3)	-0.001 (-0.2)	-0.001 (-0.4)	-0.000 (-0.2)	-0.001 (-0.2)	-0.001 (1.5)	-0.002 (-2.1)
$\Delta lnER_t$	-0.049 (-3.2)	-0.051 (-3.3)	-0.051 (-3.3)	-0.050 (-3.2)	-0.051 (-3.2)	-0.062 (-3.5)	-0.026 (-1.4)
$\Delta lnER_{t-1}$		-0.020 (-1.3)	-0.022 (-1.4)	-0.023 (-1.4)	-0.023 (-1.4)	-0.015 (-0.9)	-0.016 (-0.9)
$\Delta lnER_{t-2}$			-0.008 (-0.5)	-0.006 (-0.4)	-0.007 (-0.4)		
$\Delta lnER_{t-3}$				-0.010 (-0.6)	-0.009 (-0.6)		
$\Delta lnER_{t-4}$					-0.001 (-0.1)		
$\Delta lnCPI^*_t$	0.644 (11.2)	0.502 (5.8)	0.468 (5.1)	0.452 (4.5)	0.451 (4.5)	0.394 (4.3)	0.619 (5.9)
$\Delta lnCPI^*_{t-1}$		0.185 (2.2)	0.123 (1.3)	0.132 (1.3)	0.132 (1.2)	0.239 (2.7)	0.319 (3.0)
$\Delta lnCPI^*_{t-2}$			0.112 (1.2)	0.106 (1.1)	0.108 (1.1)		
$\Delta lnCPI^*_{t-3}$				0.010 (0.1)	0.012 (0.1)		
$\Delta lnCPI^*_{t-4}$					-0.003 (-0.1)		
$\Delta lnGDP_t$	-0.073 (-1.4)	-0.074 (-1.5)	-0.065 (-1.2)	-0.076 (-1.3)	-0.077 (-1.2)	-0.150 (-2.7)	-0.051 (-0.8)
Adj. R^2	0.489	0.507	0.506	0.500	0.49	0.504	0.377
$\beta^{IMP,SHORT}$	**-0.049**	**-0.051**	**-0.051**	**-0.050**	**-0.051**	**-0.062**	**0**
$\beta^{IMP,MEDIUM}$		**-0.090**	**-0.081**	**-0.089**	**-0.090**	**-0.077**	**0**

Appendix 4: Unit Root Tests

To deal with the problem of whether a time series follows a unit root we can apply an Augmented Dickey-Fuller (ADF) test. The results are presented in the following tables. Numbers in parenthesis are MacKinnon (1996) one-sided ρ-values. The null hypothesis will be rejected if the ρ-value is less than 5%. The lag length k is chosen by the Modified Akaike info criterion MAIC (Ng and Perron, 2001). EViews 8 chooses a maximum lag of k_{max}=13.

Variable	Level		First difference		I(d)
	k	Test value	k	Test value	
Full sample period: 1973:Q1 to 2014:Q4 (T = 168)					
Included in test equation: … intercept					
IMP	8	-2.505 (0.059)	0	-6.431 (0.000)	I(1)
CPIIMP	1	-3.604 (0.007)	8	-3.473 (0.010)	I(0)
CPI	5	-2.211 (0.203)	8	-3.279 (0.018)	I(1)
ER	1	-2.668 (0.082)	2	-6.233 (0.000)	I(1)
RER	1	-2.684 (0.079)	2	-6.663 (0.000)	I(1)
Included in test equation: … time trend und intercept					
IMP	8	-2.714 (0.232)	0	-6.452 (0.000)	I(1)
CPIIMP	2	-2.545 (0.306)	0	-9.419 (0.000)	I(1)
CPI	1	-1.154 (0.916)	5	-4.146 (0.007)	I(1)
ER	1	-3.195 (0.089)	2	-6.455 (0.000)	I(1)
RER	0	-3.368 (0.059)	0	-10.412 (0.000)	I(1)
Sample period: 1973:Q1 to 2007:Q3 (T = 139)					
Included in test equation: … intercept					
IMP	8	-2.713 (0.075)	2	-5.718 (0.000)	I(1)
CPIIMP	1	-3.101 (0.029)	8	-3.433 (0.012)	I(0)
CPI	3	-1.571 (0.494)	4	-3.162 (0.025)	I(1)
ER	1	-3.313 (0.016)	0	-8.651 (0.000)	I(0)
RER	0	-3.383 (0.015)	0	-9.161 (0.000)	I(0)
Included in test equation: … time trend und intercept					
IMP	8	-2.691 (0.242)	2	-5.629 (0.000)	I(1)
CPIIMP	1	-3.268 (0.076)	8	-3.649 (0.029)	I(1)
CPI	1	-1.290 (0.886)	4	-3.411 (0.042)	I(1)
ER	1	-2.611 (0.276)	0	-9.064 (0.000)	I(1)
RER	0	-3.040 (0.125)	0	-9.236 (0.000)	I(1)

Appendix 5: Volatility

Volatility: Effective exchange rates and price indices

(a) Variance (VAR)

Variable	full sample	sub sample periods				
	74:Q1 - 14:Q4	80:Q1 - 14:Q4	00:Q1 - 14:Q4	06:Q1 - 14:Q4	07:Q4 - 14:Q4	10:Q1 - 14:Q4
(1) Nominal effective exchange rate, 24 countries (ER)	37,2	25,4	16,6	25,2	23,4	31,6
(2) Real effective exchange rate, 24 countries (RER)	30,9	24,8	15,2	24,7	24,4	32,6
(3) Relative price: CPI/CPI*	4,8	2,2	0,3	0,7	0,8	0,7
(4) CPI	5,1	3,5	0,6	1,0	1,1	0,3
(5) Bilateral exchange rate, against the Euro	28,6	19,8	23,2	33,2	31,1	38,7
(6) Real effective exchange rate, against Germany, tradeable goods				33,7	31,7	39,4
(7) CPI_T/CPI*_T (tradeable goods)				0,9	1,0	1,1
(8) CPI_T (tradeable goods)			1,0	1,3	1,5	0,4
Calculations						
(9) Covarianz [(2)-(1)-(3)]/2	-5,5	-1,4	-0,9	-0,6	0,1	0,1
(10) Covarianz [(6)-(5)-(7)]/2				-0,2	-0,2	-0,2
(11) Statistical ERPT-coeffizient beta (9)/(1)	-0,15	-0,05	-0,05	-0,02	0,01	0,00
(11) Statistical ERPT-coeffizient beta (10)/(5)				-0,01	-0,01	-0,01

(b) Mean Squared Error (MSE)

Variable	full sample	sub sample periods				
	74:Q1 - 14:Q4	80:Q1 - 14:Q4	00:Q1 - 14:Q4	06:Q1 - 14:Q4	07:Q4 - 14:Q4	10:Q1 - 14:Q4
(1) Nominal effective exchange rate, 24 countries (ER)	44,8	28,3	24,6	33,3	39,5	46,8
(2) Real effective exchange rate, 24 countries (RER)	31,6	24,9	19,4	26,6	30,6	36,6
(3) Relative price: CPI/CPI*	8,7	4,0	2,3	2,8	3,2	4,3
(4) CPI	9,6	6,9	1,1	1,1	1,2	0,3
(5) Bilateral exchange rate, against the Euro	31,2	21,0	28,7	40,7	48,0	58,5
(6) Real effective exchange rate, against Germany, tradeable goods				35,1	37,8	46,1
(7) CPI_T/CPI*_T (tradeable goods)				3,3	3,4	4,3
(8) CPI_T (tradeable goods)			1,2	1,4	1,5	0,4

Raw data for calculations: first-order differences of logarithm values, quarterly data.

1. Engel (1999) uses the mean squared error (MSE) as volatility measure. The MSE of the change of the respective indices is the sum of the variance and square of the arithmetic mean.

2. See Appendix 1 for approximation of tradable goods.

Source: SNB, FSO, Federal Statistic Office of Germany (Destatis), author's calculations.

Prof. Dr. Kersten Kellermann

Kersten Kellermann studied economics at the Universities of Konstanz and Western Ontario (Canada). In 1996, she received her doctorate at the University of Cologne on the subject „The interregional convergence of labour productivity". In 2008, the venia legendi was awarded to her by the faculty of economics and social science of the University of Fribourg in Switzerland. She has been professor for economics and economic policy at Deggendorf Institute of Technology since 2015. Kersten Kellermann`s research is focused on evidence-based economic policy and regional economic growth.

Kersten Kellermann studierte Volkswirtschaftslehre an den Universitäten Konstanz und Western Ontario (Kanada). 1996 promovierte sie an der Universität Köln mit einer Arbeit zum Einfluss des Finanzausgleichs auf die Konvergenz von Regionen. 2008 erhielt sie die Venia Legendi für Volkswirtschaftslehre von der Wirtschafts- und Sozialwissenschaftlichen Fakultät der Universität Fribourg (Schweiz). Im Dezember 2014 wurde Frau Kellermann auf die Professur für Wirtschaftspolitik an die Technische Hochschule Deggendorf berufen. Sie ist Senior Research Fellow der Konjunkturforschungsstelle Vierländereck (KOVL).

Kontakt / *Contact*

kersten.kellermann@th-deg.de

Dr. Carsten-Henning Schlag

Carsten-Henning Schlag studied economics at the University of Kiel, with a focus on econometrics, public finance, and macroeconomics. In 1998, he received his doctorate at the University of Cologne on the subject „Public infrastructure and economic growth in Germany". He worked as a Senior Economist for the German Council of Economic Experts in Wiesbaden, and for the Swiss Institute of Economics (KOF) at the ETH Zurich. From 2004-2014 he was Head of the Liechtenstein Institute of Economics in Vaduz. Carsten-Henning Schlag is currently the managing director of the Konjunkturforschungsstelle Vierländereck (KOVL) and was a DAAD Visiting Professor at Deggendorf Institute of Technology during the summer semester of 2016.

Carsten-Henning Schlag studierte Volkswirtschaftslehre an der Universität Kiel, mit Schwerpunkt Ökonometrie, Finanzwissenschaft und Makroökonomik. 1998 promovierte er an der Universität Köln mit einer ökonometrischen Arbeit zur Öffentlichen Kapital-Hypothese. Danach war Herr Schlag im wissenschaftlichen Stab des Sachverständigenrates zur Begutachtung der gesamtwirtschaftlichen Entwicklung in Wiesbaden tätig und wechselte anschließend als Senior Economist an die Konjunkturforschungsstelle (KOF) der Eidgenössischen Technischen Hochschule (ETH) Zürich. Im Mai 2004 übernahm Herr Schlag die Aufgabe, eine Konjunkturforschungsstelle nach Schweizer Vorbild für das Fürstentum Liechtenstein aufzubauen und lenkte deren Geschicke in den folgenden zehn Jahren. Er ist Geschäftsführer des volkswirtschaftlichen Forschungs- und Beratungsinstituts Konjunkturforschungsstelle Vierländereck (KOVL) und hatte im Sommersemester 2016 eine DAAD-Gastprofessur an der TH Deggendorf inne.

Kontakt / *Contact*

✉ **carsten.schlag@kovl.li**

BAVARIAN JOURNAL OF APPLIED SCIENCES®

Die Entwicklung der Langfristzinsen in den USA und das *Quantitative Easing* der FED*

Harm Bandholz
UniCredit Bank AG

Franz Seitz
Ostbayerische Technische Hochschule Amberg-Weiden

Jörg Clostermann
Technische Hochschule Ingolstadt

ABSTRACT

Seit der Finanzkrise verfolgt die US-amerikanische Notenbank (FED) einen sehr expansiven geldpolitischen Kurs. Die Zinsen wurden auf historische Niedrigstniveaus gesenkt, darüber hinaus griff die FED auf verschiedene Maßnahmen quantitativer Lockerungen zurück. Trotz dieser Ereignisse zeigen die ökonometrischen Analysen in diesem Papier, dass die klassischen Einflussfaktoren (Notenbankzinsen, Inflationserwartungen, Auslandsnachfrage nach US-Staatsanleihen) weiterhin signifikante Erklärungskraft besitzen. Allerdings sind diese Zusammenhänge in den letzten Jahren instabiler geworden. Darüber hinaus gibt es Anzeichen dafür, dass das Anleihenaufkaufprogramm der FED einen messbaren Einfluss auf die Langfristzinsen hat.

JEL Klassifikationsnummern: E43, E52, E58.

Since the financial crisis, the FED has followed an expansionary monetary policy stance. Interest rates have reached historical low levels and the FED introduced several variants of quantitative easing. Despite these actions, our econometric results show that traditional determinants of long-term interest rates (central bank rates, inflationary expectations, foreign demand for US bonds, the business cycle) still exert a significant influence. However, it turned out that the relationships have become more unstable in the last few years. Moreover, the asset purchases of the FED seem to have contributed to the low levels of long-term rates, at least partly.

KEYWORDS

Geldpolitik, USA, Finanzkrise, Staatsanleihenrendite, Federal Reserve Board, Quantitative Easing, Zinsprognose, Fehlerkorrekturmodell

* Wir danken einem anonymen Gutachter für wertvolle Hinweise und Verbesserungsvorschläge.

Monetary policy, USA, financial crisis, government bond yield, FED, quantitative easing, interest rate forecast, error correction model

1. Einführung

Mit der Finanzkrise ist die US-amerikanische Notenbank Federal Reserve (FED) auf einen sehr expansiven geldpolitischen Kurs eingeschwenkt. Das Federal Funds Rate Target, die Zinsrate, zu der US-Banken untereinander Geld leihen, wurde bereits 2009 auf nahe 0% gesenkt; darüber hinaus griff die FED auf verschiedene Maßnahmen quantitativer Lockerungen (quantitative easing) zurück.[1]

Schon vor der Finanzmarktkrise waren die US-Langfristzinsen aus historischer Perspektive betrachtet sehr niedrig. Dies veranlasste Alan Greenspan, den damaligen Chairman der FED, im Februar 2005 von einem „Conundrum" zu sprechen.[2] Viele „außergewöhnliche Ursachen" für diese Entwicklung wurden in der Folgezeit ins Feld geführt.[3] Trotz der vermeintlich aus der Zinsperspektive betrachteten „schwer erklärlichen" Situation zeigen Bandholz et. al. 2009 [1], dass man mit wenigen klassischen Determinanten (Kurzfristzins, Inflationserwartungen, Konjunktur, Auslandsnachfrage nach US-Staatspapieren) die Entwicklung der zehnjährigen Treasury Yields (US-Staatsanleiherenditen) während der Greenspan Ära (1985-2005) sehr gut erklären kann und daraus abgeleitete Zinsprognosen besser als Random-Walk-Modelle[4] sind.

Vor diesem Hintergrund scheint es angebracht, das Zinsmodell von Bandholz et. al. [1] als Referenzmodell zu definieren und zu prüfen, ob es weiterhin, d. h. auch in Zeiten des quantitative easing und der Nullzinspolitik der FED, geeignet ist, die Entwicklung zu erklären und gute Prognosen für die zehnjährigen US-Staatsanleiherenditen zu generieren.

Die weiteren Ausführungen gliedern sich wie folgt: In einem ersten Schritt wird die zugrundeliegende Theorie des Modells von Bandholz et. al. [1] nachgezeichnet und das daraus folgende empirische Modell geschätzt, einmal für den von Bandholz et. al. [1] gewählten, dann für den bis zum aktuellen Rand verlängerten Zeitraum. Nach einem Vergleich der Ergebnisse erweitern wir das Modell um eine weitere Variable, die monetäre Basis. Diese Variable steht für die massive Ausweitung der Zentralbankbilanz infolge verschiedener geldpolitischer Maßnahmen der FED als Reaktion auf die Finanzkrise und deren Folgewirkungen. Auch dieses Modell wird interpretiert und seine Erklärungskraft und Prognoseperformance evaluiert. Die Arbeit endet mit einer Zusammenfassung der Ergebnisse und Schlussfolgerungen.

2. Das theoretische Modell und die Daten

Wie viele Zinsanalysen von Banken, Zentralbanken und internationalen Organisationen wie IWF, OECD oder BIZ stellen auch Bandholz et. al. [1] bezüglich der Entwicklung des US-Langfristzinses u. a. auf folgende drei Faktoren ab: Die Inflationserwartungen, den Konjunkturverlauf und die Zinspolitik der US-Notenbank. Theoretisch lässt sich die Wahl dieser drei Faktoren wie folgt begründen:

Nach der Fisher-Gleichung[5] kann man den Nominalzins in den Realzins und die erwartete Inflationsrate zerlegen. Sind die Realzinsen stationär, schlagen sich die Inflationserwartungen direkt proportional in Schwankungen des Nominalzinses nieder.

Nach der „Loanable-Funds"-Theorie[6] erhöht sich bei einem Konjunkturaufschwung das Angebot sowie die Nachfrage nach langfristigen Wertpapieren. Zusätzlich sinkt mit einem verbesserten konjunkturellen Umfeld die Risikoprämie dieser Wertpapiere. Ist der Nachfrageeffekt nur schwach ausgeprägt, schwanken die Renditen positiv mit der Konjunktur.

Nach der Erwartungs- und der

[1] Eine ausführliche Beschreibung dieser geldpolitischen Maßnahmen findet sich bei Belke et. al. [2] und Fawley/Neely [3].
[2] Greenspan, 2005 [4].
[3] Bernanke, 2006 [5]. Eine intensive Analyse für das seit 30 Jahren zu beobachtenden Absinken der langfristigen US-Zinsen findet sich in Obstfeld/Tesar [6] und The Executive Office of the President of the United States [7].
[4] Auer/Rottmann, 2015 (S. 587 ff.) [8].
[5] Issing, 2011 (S. 258) [9]
[6] Issing, 2011 (S. 188 ff.) [9].

Liquiditätsprämientheorie[7] der Zinsstruktur besteht ein direkt proportionaler Zusammenhang zwischen dem aktuellen kurzfristigen und dem langfristigen Zins. Zudem postulieren diese Theorien, dass die langfristigen Zinsen den zukünftigen Pfad der kurzfristigen Zinsen widerspiegeln (gegeben die Laufzeitenprämie)[8]. Demnach würde man einen kausalen Zusammenhang vom aktuellen langfristigen zum aktuellen kurzfristigen Zins vermuten. Allerdings hat die Notenbank eine dominante Stellung am kurzfristigen Ende des Marktes. Jede Verlautbarung und Aktion wird von den Märkten geprüft und hinsichtlich der zukünftigen Entwicklung der (kurzfristigen) Zinsen bewertet. Demnach ist es realistisch, anzunehmen, dass die kurzfristigen Zinsen auch die langfristigen Zinsen beeinflussen.

Neben diesen drei klassischen Zinsbestimmungsfaktoren inkludieren Bandholz et. al. [1] noch eine weitere Variable, welche die strukturelle Erhöhung der Auslandsnachfrage nach US-Staatspapieren einfangen soll.[9]

Bei den ökonometrischen Schätzungen in den nächsten Kapiteln greifen wir auf folgende monatlichen US-Daten zurück:[10] Die zehnjährige US-Rendite (=LZINS), den dreimonatigen US-Geldmarktzins (=KZINS), als Maß für die Inflationserwartungen die Jahreswachstumsrate der Core Inflation (Headline Consumer Price Index ohne Nahrungsmittel und Energie) (=CORE), als Konjunkturmaß den ISM Index des verarbeitenden Gewerbes vom Institute for Supply Management (=ISM) und als Auslandsnachfrage (=HOLDING) auf den Quotienten von „Rest of the World Treasury Securities“ zu „Federal Government Debt Securities“.[11] In Schaubild 1 sind die Variablen dargestellt.

[7] Issing, 2011 (S. 123 ff.) [9].
[8] Issing, 2011 (S. 123-124) [9].
[9] Bandholz et. al. [1] vermuten, dass die Erhöhung von Devisenreserven einiger asiatischer Länder, das Recycling der Petro-Dollar, das veränderte Anlageverhalten von institutionellen Investoren und die weltweite Erhöhung der Liquidität die Risikoprämie für US-amerikanische Staatspapiere und infolgedessen auch das Zinsniveau reduziert hat.
[10] Die Daten können von den Autoren angefordert werden.
[11] Board of Governors of the Federal Reserve System [10], L132, Zeile 10 und L106 Zeile 20, jeweils von Quartalszahlen in Monatsdaten umgerechnet.

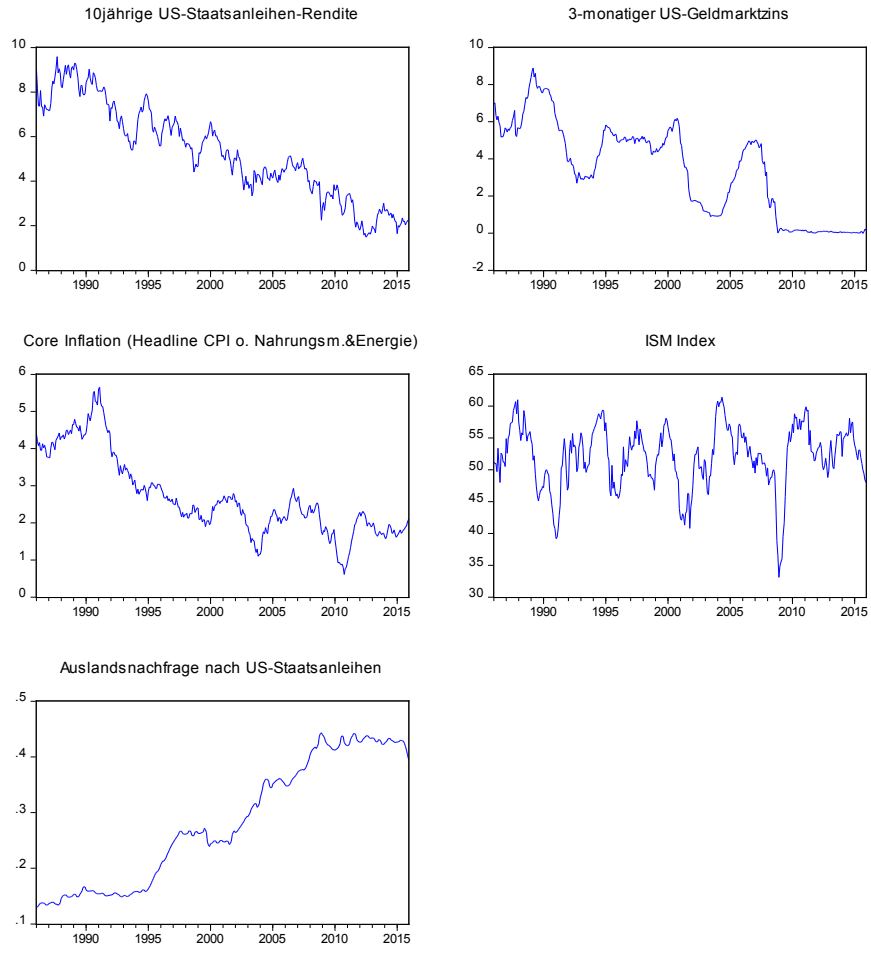


Schaubild 1: Die Variablen des BCS-Modells

3. Ökonometrische Ergebnisse

Nach einer Vielzahl von Spezifikationstests präsentierten Bandholz et. al. [1] als „Best Of" ein Einzelgleichungsfehlerkorrekturmodell. Schätzt man dieses Modell nach (im folgenden BCS-Modell genannt), so erhält man für den gleichen, damals zugrunde gelegten Schätzzeitraum:

$$\Delta LZINS_t = \underset{(5,7)}{-0.26} \cdot (LZINS_{t-1} + \underset{(2.6)}{9.99} - \underset{(6.6)}{0.31} \cdot KZINS_{t-1} - \underset{(5.2)}{0.65} \cdot CORE_{t-1} + \underset{(3.4)}{5.87} \cdot HOLDING_{t-1})$$
$$+ \underset{(2.2)}{0.14} \cdot \Delta LZINS_{t-1} + \underset{(7.1)}{0.55} \cdot \Delta KZINS_t - \underset{(2.4)}{0.21} \cdot \Delta KZINS_{t-1} + \underset{(3.8)}{1.91} \cdot \ln(ISM_t) - \underset{(2.1)}{0.99} \cdot \ln(ISM_{t-1})$$

$R^2 = 0.32$; Adj. $R^2 = 0.30$; Standardfehler = 0.25; F-Wert= 11,98.

Abbildung 1: Schätzergebnisse für das BCS-Modell 1986.01-2005.09 (t-Werte in Klammern)

Auf den ersten Blick fällt auf, dass die Koeffizienten in Abbildung 1 nicht vollkommen identisch mit den Koeffizienten in der Originalpublikation von Bandholz et. al. [1] sind. Verantwortlich für die leichten Abweichungen dürften Datenrevisionen sein.

Verlängert man den Schätzzeitraum bis Dezember 2015 (2015.12), erhält man folgende Schätzergebnisse:

$$\Delta LZINS_t = \underset{(5.2)}{-0.14} \cdot (LZINS_{t-1} + \underset{(0.2)}{0.76} - \underset{(4.7)}{0.33} \cdot KZINS_{t-1} - \underset{(2.3)}{0.40} \cdot CORE_{t-1} + \underset{(4.8)}{8.37} \cdot HOLDING_{t-1})$$
$$+ \underset{(1.4)}{0.07} \cdot \Delta LZINS_{t-1} + \underset{(7.5)}{0.47} \cdot \Delta KZINS_t - \underset{(2.1)}{0.14} \cdot \Delta KZINS_{t-1} + \underset{(4.4)}{1.72} \cdot \ln(ISM_t) - \underset{(4.0)}{1.51} \cdot \ln(ISM_{t-1})$$

$R^2 = 0.25$; Adj. $R^2 = 0.23$; Standardfehler = 0.25; F-Wert= 12,78.

Abbildung 2: Schätzergebnisse für das BCS-Modell 1986.01-2015.12 (t-Werte in Klammern)

Folgende Dinge sind auffällig:

- Der Anteil der erklärten Streuung fällt von 30% auf 25%.
- Abweichungen von der langfristigen Kointegrationsbeziehung werden nun langsamer abgebaut. Der Koeffizient des Fehlerkorrekturterms sinkt von 0,26 auf 0,14.
- Die langfristige Semi-Elastizität der Kerninflation ist deutlich von 0,65 auf 0,40 gesunken.
- Die langfristige Reaktion des Langfristzinses auf Veränderungen der Auslandsnachfrage ist dagegen elastischer geworden (von -5,87 auf -8,37).

Führt man einen Wald-Test auf Koeffizientengleichheit insgesamt durch, muss diese Nullhypothese mit einer Irrtumswahrscheinlichkeit von nahezu 0% abgelehnt werden (siehe Abbildung 3).

Test Statistic	Value	df	Probability
F-statistic	4.160772	(10, 350)	0.0000
Chi-square	41.60772	10	0.0000

Abbildung 3: Wald-Test auf Koeffizientengleichheit

Ein Blick auf die Residuen der Schätzgleichung aus Abbildung 2 erweckt allerdings nicht den Eindruck, dass sich seit der amerikanischen Finanzkrise etwas Grundlegendes geändert hat (siehe Schaubild 2). Die Ausschläge am aktuellen Rand halten sich im historischen Rahmen. Auch das Muster scheint sich nicht auffällig geändert zu haben.

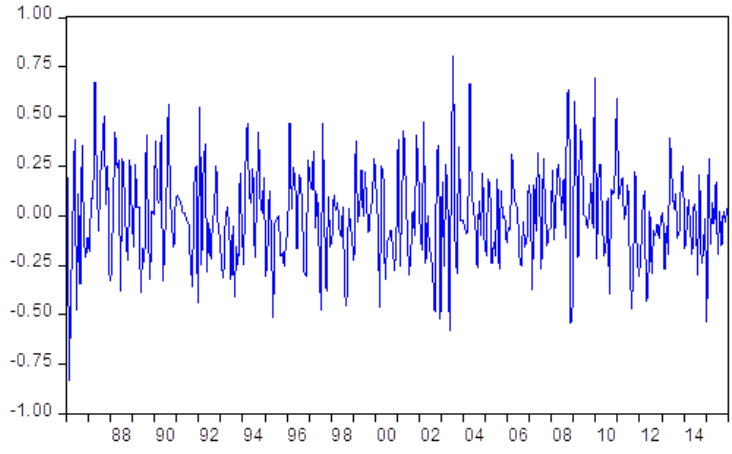


Schaubild 2: Residuen des BCS-Modells mit erweitertem Schätzzeitraum

Der Cusum-Test und der Cusum-Square-Test (Irrtumswahrscheinlichkeit < 10%) geben keine Hinweise auf Instabilitäten des Schätzansatzes. Auf der anderen Seite zeigt der Chow-Strukturbruch-Test (siehe Abbildung 4), dass die Nullhypothese keines Strukturbruchs ab Oktober 2005 (2005.10) abgelehnt werden muss.

Chow Breakpoint Test: 2005M10
Null Hypothesis: No breaks at specified breakpoints
Varying regressors: All equation variables
Equation Sample: 1986M01 2015M12

F-statistic	2.785799	Prob. F(10,340)	0.0025
Log likelihood ratio	28.35049	Prob. Chi-Square(10)	0.0016
Wald Statistic	27.85799	Prob. Chi-Square(10)	0.0019

Abbildung 4: Chow Breakpoint Test

Bandholz et. al. [1] führen in ihrem Papier auch umfangreiche Out-of-Sample-Tests durch, d. h. sie prognostizieren den US-Zins über den zugrunde liegenden Stützzeitraum hinaus. Sie generieren 120 rekursive Vorhersagen bis zu 12 Monate im Voraus, wobei sie die unabhängigen Variablen jeweils auf Basis eines für die jeweilige Variable angepassten Random-Walk-Modells prognostizieren.[12] Als Prognosekriterien dienen der sogenannte Root-Mean-Squared-Error (RMSE), der die durchschnittliche Prognoseabweichung misst, und die Vorzeichentrefferquote, die angibt, wie hoch der Anteil der korrekten Richtungsvorhersage ist. Vergleicht man anhand dieser beiden Prognosegütemaße die Prognosen für den zehnjährigen US-Staatsanleihen, sieht man, dass die Vorhersagequalität des BCS-Modells auch für den bis Dezember 2015 erweiterten Stützzeitraum ähnlich gut ist (s. Abbildung 5).

	Prognoseprozedure: Fully Dynamic			
Prognosezeitraum	RMSE in Basispunkte		Vorzeichentrefferquote in %	
... Monate voraus	Bandholz et.al.	Neuer Zeitraum	Bandholz et.al.	Neuer Zeitraum
1	26.76	25.11	54.17	55.00
2	38.40	36.63	55.00	60.00
3	45.16	43.53	58.33	60.83
4	51.46	50.25	61.67	60.83
5	55.48	56.07	56.67	60.00
6	57.12	60.32	60.00	55.83
7	58.60	63.07	65.00	59.17
8	60.76	64.55	60.00	60.83
9	62.71	65.18	67.50	65.00
10	65.72	65.95	69.17	65.83
11	68.38	67.38	66.67	68.33
12	71.49	69.15	70.00	69.17

Abbildung 5: Vergleich von RMSE und Vorzeichentreffer des BCS-Modells

4. Die Erweiterung um die monetäre Basis

Wie eingangs erwähnt, verfolgt die FED seit der Finanzkrise einen sehr expansiven geldpolitischen Kurs.[13] Die Notenbankzinsen wurden kräftig gesenkt, darüber hinaus griff die FED auf verschiedene Maßnahmen quantitativer Lockerungen zurück. Ziel der quantitativen Lockerung ist es, die Kapitalmarktzinsen auf breiter Front zu drücken und damit das gesamte Zinsspektrum zu beeinflussen, um letztendlich Kreditnachfrage und Güternachfrage anzukurbeln. Als Maß für die quantitative Lockerung verwenden wir die monetäre Basis. Setzt man diese Größe in Relation zum nominalen Bruttoinlandsprodukt der USA, so sieht man in Schaubild 3, dass seit dem Ausbruch der Finanzkrise 2008 dieses Verhältnis kräftig gestiegen ist.

[12] Bandholz et al., S. 547 ff. [1].
[13] Dieser wurde durch die sukzessive Rückführung der Wertpapierkäufe seit Oktober 2014 und die im Dezember 2015 erfolgte geringfügige Zinserhöhung um 0,25 Prozentpunkte im Zeitablauf abgeschwächt.

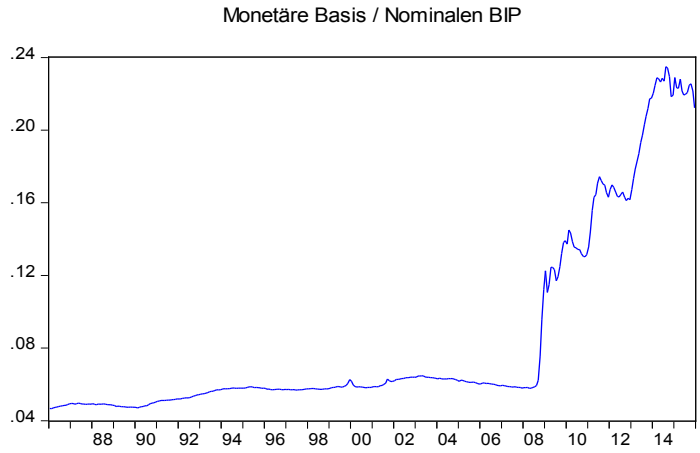


Schaubild 3: Monetäre Basis im Verhältnis zu nominalen BIP

Im nächsten Schritt versuchen wir, diese Variable in den Schätzansatz zu integrieren, um beurteilen zu können, ob und wie stark eine Ausweitung der monetären Basis auf den zehnjährigen US-Zins wirkt.[14]

Wir wählen dabei das von Bandholz et. al. [1] beschriebene Prozedere, d. h. wir schätzen ein Vektorfehlerkorrekturmodell (=VECM) und prüfen, ob es in einen Einzelgleichungsfehlerkorrekturansatz (=SEECM) überführbar ist.

Bis auf den ISM Index des verarbeitenden Gewerbes vom Institute for Supply Management (=ISM) sind alle Variablen - die zehnjährige Rendite von US-Staatsanleihen (=LZINS), der dreimonatige US-Geldmarktzins (=KZINS), Core Inflation (Headline CPI ohne Nahrungsmittel und Energie) (=CORE), die Auslandsnachfrage nach amerikanischen Staatspapieren (=HOLDING) und das Verhältnis zwischen der monetären Basis und dem nominalen BIP (=K_MB_GDP) – scheinbar integriert vom Grade 1 bzw. nicht stationär.[15] Damit ist die Grundvoraussetzung für die Anwendung der Kointegrationsanalyse erfüllt. Im nächsten Schritt wird die Lagordnung bestimmt. Dafür wird ein unrestringiertes VAR-Modell nur mit den I(1)-Variablen geschätzt. Die Informationskriterien von Schwarz und Hannan-Quinn (siehe Abbildung 6) empfehlen eine optimale Laglänge von 2.

Lag	Schwartz	Hannan-Quinn
0	1.5549	1.5224
1	-18.7031	-18.8982
2	-19.0260	-19.3836
3	-18.7454	-19.2656
4	-18.6037	-19.2865
5	-18.3245	-19.1698
6	-18.0449	-19.0528
7	-17.7906	-18.9611
8	-17.6525	-18.9855

Abbildung 6: Test auf Laglänge

Die Anzahl der Kointegrationsbeziehungen zwischen den I(1)-Variablen wird mit Hilfe der Trace- und der Maximum-Eigenvalue-Statistik bestimmt. Beide Tests empfehlen eine Restriktion des Systems auf nur eine Kointegrationsbeziehung (Irrtumswahrscheinlichkeit < 1%, siehe Abbildung 7).

[14] Die Stärke des Einflusses der monetären Basis auf den zehnjährigen US-Zins hängt auch von der Laufzeitenstruktur der Staatspapiere ab, welche die FED aufgekauft hat. Allerdings sind hierüber keine öffentlichen Daten verfügbar.
[15] Im Anhang 1 und 2 finden sich die dazugehörigen Unit-Root- bzw. Stationaritätstests.

Unrestricted Cointegration Rank Test (Trace)				
Hypothesized		Trace	0.05	
No. of CE(s)	Eigenvalue	Statistic	Critical Value	Prob.**
None *	0.12	84.82	69.82	0.00
At most 1	0.05	37.99	47.86	0.30
At most 2	0.03	20.20	29.80	0.41
At most 3	0.01	8.47	15.49	0.42
At most 4	0.01	3.67	3.84	0.06
Trace test indicates 1 cointegrating eqn(s) at the 0.05 level				
* denotes rejection of the hypothesis at the 0.05 level				
**MacKinnon-Haug-Michelis (1999) p-values				

Unrestricted Cointegration Rank Test (Maximum Eigenvalue)				
Hypothesized		Max-Eigen	0.05	
No. of CE(s)	Eigenvalue	Statistic	Critical Value	Prob.**
None *	0.12	46.82	33.88	0.00
At most 1	0.05	17.79	27.58	0.51
At most 2	0.03	11.73	21.13	0.57
At most 3	0.01	4.80	14.26	0.77
At most 4	0.01	3.67	3.84	0.06
Max-eigenvalue test indicates 1 cointegrating eqn(s) at the 0.05 level				
* denotes rejection of the hypothesis at the 0.05 level				
**MacKinnon-Haug-Michelis (1999) p-values				

Abbildung 7: Test auf Anzahl der Kointegrationsbeziehungen

Kointegrationsgleichung	
$LZINS_{-1}$	1.00000
$KZINS_{-1}$	-0.26525
	[-4.35713]
$HOLDING_{-1}$	8.21711
	[5.41273]
$CORE_{-1}$	-0.48026
	[-3.41231]
$K\_MB\_GDP_{-1}$	5.71388
	[2.42409]
Konstante	-5.89297

Kurzfristgleichung	ΔLZINS	ΔKZINS	ΔHOLDING	ΔCORE	ΔK_MB_GDP
Fehlerkorrekturen	-0.15605	0.00073	0.00019	0.01904	0.00077
	[-5.28019]	[0.03058]	[-0.68799]	[1.35029]	[2.84953]
Δ $LZINS_{-1}$	0.10647	0.06238	0.00054	-0.02321	-0.00040
	[1.94411]	[1.40579]	[1.05008]	[-0.88820]	[-0.80238]
Δ $KZINS_{-1}$	-0.13628	0.02009	-0.00157	0.02883	-0.00056
	[-1.91391]	[0.34822]	[-2.34103]	[0.84853]	[-0.87094]
Δ $HOLDING_{-1}$	-9.06862	-6.57977	0.61092	-0.81610	0.05090
	[-2.01327]	[-1.80275]	[14.4159]	[-0.37969]	[1.24242]
Δ $CORE_{-1}$	0.15094	-0.04981	-0.00081	0.11323	-0.00079
	[1.31204]	[-0.55392]	[-0.77244]	[2.13826]	[-0.77767]
Δ $K\_MB\_GDP_{-1}$	15.89592	5.21385	-0.00028	-0.30317	0.41374
	[2.88721]	[1.16873]	[-0.00539]	[-0.11540]	[8.26323]
Konstante	-2.82349	-2.94354	-0.01289	-0.32415	0.01316
	[-3.96326]	[-5.09916]	[-1.92281]	[-0.95355]	[2.03152]
LN(ISM)	2.28073	1.01341	-0.00698	-0.03716	-0.00619
	[5.59923]	[3.07044]	[-1.82137]	[-0.19118]	[-1.67110]
$Ln(ISM_{-1})$	-1.57025	-0.27165	0.01031	0.11811	0.00291
	[-3.93317]	[-0.83976]	[2.74399]	[0.62002]	[0.80028]
R^2	0.15786	0.14289	0.41205	0.03741	0.25231
Adj. R^2	0.13867	0.12336	0.39865	0.01547	0.23526
Standardfehler	0.26547	0.21511	0.00250	0.12670	0.00241
F-Wert	8.22450	7.31448	30.74834	1.70510	14.80531

Abbildung 8: VECM (t-Werte in Klammern)

Bei der Schätzung des VECM berücksichtigen wir in Analogie zu Bandholz et. al. [1] den ISM als Konjunkturvariable (logarithmiert) kontemporär und mit der ersten Verzögerung als stationäre exogene Variable. Die Schätzergebnisse (Abbildung 8) zeigen, dass alle Langfristvariablen signifikant sind und das richtige Vorzeichen aufweisen. In diesem Zusammenhang fällt auf, dass die monetäre Basis einen negativen Effekt auf den Zins hat. Dieser Effekt wird auch durch andere Untersuchungen bestätigt.[16] Der Koeffizient des Fehlerkorrekturterms der LZINS-Gleichung ist hoch signifikant negativ (siehe Abbildung 8) und deutet auf die (schwache) Endogenität der zehnjährigen US-Staatsanleiherendite hin. Restringiert man das System in der Art, das man (schwache) Exogenität aller anderen Variablen außer LZINS unterstellt, kann man diese Hypothese nur mit einer Irrtumswahrscheinlichkeit von 4,5% nicht ablehnen (Abbildung 9).[17]

Cointegration Restrictions:	
A(2,1)=0, A(3,1)=0, A(4,1)=0, A(5,1)=0	
Convergence achieved after 4 iterations.	
Not all cointegrating vectors are identified	
LR test for binding restrictions (rank = 1):	
Chi-square(4)	9.749467
Probability	0.044866

Abbildung 9: Test auf (schwache) Exogenität im VECM

Der Befund einer schwach endogenen Staatsanleihenrendite erlaubt es, auf ein Einzelgleichung-Fehlerkorrekturmodell (=SEECM) überzugehen[18], wodurch sich die Effizienz der Schätzung verbessern kann. Das SEECM entspricht der Zinsgleichung des VECM, nur dass für alle stationären Variablen der Kurzfristgleichung kontemporäre und, wie von Banerjee et. al. [12] empfohlen, u. a. bis zu drei Perioden verzögerte Werte aufgenommen werden.[19] In einem nachfolgenden Auswahlprozess werden die nicht signifikanten Koeffizienten der stationären Variablen (Irrtumswahrscheinlichkeit > 1%) sukzessive auf Null gesetzt.[20] Als finales Ergebnis erhalten wir Abbildung 10.

$$\Delta LZINS_t = \underset{(5.0)}{-0.14} \cdot (LZINS_{t-1} - \underset{(3.5)}{0.28} \cdot KZINS_{t-1} - \underset{(2.7)}{0.50} \cdot CORE_{t-1} + \underset{(3.4)}{6.65} \cdot HOLDING_{t-1} + \underset{(1.5)}{4.68} \cdot K\_MB\_GDP_{t-1}) + \underset{(0.6)}{0.39} - \underset{(2.9)}{0.14} \cdot \Delta LZINS_{t-2} + \underset{(8.2)}{0.50} \cdot \Delta KZINS_t + \underset{(4.6)}{1.07} \cdot \ln(ISM_t) - \underset{(4.7)}{0.99} \cdot \ln(ISM_{t-3})$$

$R^2 = 0.26$; Adj. $R^2 = 0.25$; Standardfehler = 0.25; F-Wert= 13,98; LM(1) = 2,22; LM(4) = 1.09; ARCH(1) = 0.66; ARCH(4) = 1.29; JB = 5,53; CUSUM: stabil; CUSUM Square: stabil.

Abbildung 10: SEECM mit Monetärer Basis (t-Werte in Klammern)

Die Koeffizienten der Langfristbeziehung bzw. Kointegrationsbeziehung haben das erwartete Vorzeichen und sind bis auf den Koeffizienten der monetären Basis hoch signifikant. Die geschätzten Koeffizienten gleichen im Großen und Ganzen den Ergebnissen der Langfristgleichung, die sich aus dem Johansen-Verfahren [14] (siehe Abbildung 7) ableiten lassen, und den Ergebnissen bei Bandholz et. al. [1]. Dies lässt auf eine gewisse methodenunabhängige Stabilität des Zusammenhangs schließen.

Der Koeffizient des kurzfristigen Zinses in der Kointegrationsbeziehung beträgt 0,28. Ein Anstieg des Kurzfristzinses um zehn Basispunkte würde den zehnjährigen Zins um ca. drei Basispunkte ansteigen lassen. Die Zinsstrukturkurve würde in diesem Fall bzw. bei einer Zinsstraffungspolitik der Notenbank flacher (oder sogar invers) werden.

Der Koeffizient der Kerninflation als Maß für die Inflationserwartungen in der Langfristbeziehung beträgt 0,5. Ein Anstieg dieser Variable um 1% würde den zehnjährigen Zins also um 50 Basispunkte ansteigen lassen.

Die Auslandsnachfrage nach US-Staatsanleihen

[16] Wu, 2014 [11]. Wu zeigt darüber hinaus, dass das Anleihenaufkaufprogramm der FED über mehrere Kanäle die Staatsanleiherenditen beeinflusst.

[17] Im Anhang 3 befinden sich die Ergebnisse des VECM mit den inkludierten Restriktionen auf schwache Exogenität.

[18] Engle et. al., 1983 [13]; Johansen, 1992 [14].

[19] Banerjee et al. [12] (S. 275) empfehlen, das SEECM auch durch Leads der stationären Variablen anzureichern, um Stichprobenfehler in zu kleinen Stichproben zu reduzieren. Wir verzichten auf dieses Vorgehen, zum einen, weil die Stichprobe mit n=360 relativ groß ist, zum anderen, weil wir später mit diesem Ansatz Out-of-Sample-Prognosen generieren möchten und uns dies nötigen würde, Lead-Variablen zu prognostizieren.

[20] Derksen/Keselman [15].

hat sich seit Mitte der achtziger Jahre bis zum heutigen Zeitpunkt fast vervierfacht (von 12% auf 40%). Bringt man diesen Sachverhalt mit den Schätzergebnissen in Verbindung, bedeutet dies für sich genommen allein schon ein Absinken der Rendite für zehnjährige US-Staatsanleihen um rund 190 Basispunkte [(0,4-0,12) ·-6,65=-1,86].[21]

Der Koeffizient des Verhältnisses von monetärer Basis und dem nominalen US-BIP beträgt -4,7. Dieser Quotient stieg zwischen 2008 und 2015 von 6% auf 22%. Dieser starke Anstieg impliziert eine langfristige Senkung der Staatsanleihenrendite um ca. 75 Basispunkte [(0,22-0,06) ·-4,7=0,75].[22] Dieses Ergebnis ist mit Vorsicht zu interpretieren, da der Koeffizient mit einer Irrtumswahrscheinlichkeit von 12,3% nur schwach abgesichert ist.

Eine Erhöhung des ISM-Index erhöht den zehnjährigen Zins. Konjunkturänderungserwartungen haben allerdings nur in der sehr kurzen Frist eine hohe Bedeutung, da dieser Effekt in den darauf folgenden drei Monaten wieder schnell ausebbt. Man kann deshalb von einer Art Überschießen des Zinses auf Konjunkturerwartungen sprechen.

Der Koeffizient des Fehlerkorrekturterms ist negativ und gemessen an den kritischen Werten von Banerjee et al. [12] signifikant von null verschieden. Damit ist die Bedingung für ein langfristiges stabiles Gleichgewicht bzw. eine Kointegrationsbeziehung zwischen den I(1)-Variablen erfüllt. Die Anpassungsgeschwindigkeit der Staatsanleihenrendite an seinen fundamental bestimmten Gleichgewichtswert mit einem Parameterwert von -0,14 kann mit einer Halbwertszeit von vier bis fünf Monaten gleichgesetzt werden, d. h. dass die Staatsanleihenrendite nach einem exogenen Schock in weniger als fünf Monaten den Abstand zu seinem Gleichgewichtskurs zur Hälfte abbaut.

Bei einer Irrtumswahrscheinlichkeit von < 5% geben Breusch-Godfrey-Lagrange-Multiplikator-Tests (LM) keine Hinweise auf Autokorrelation in den Residuen (1. und 4. Ordnung), desgleichen kann der Lagrange-Multiplikator-Test auf autoregressiv bedingte Heteroskedastizität (1. und 4. Ordnung) in den Störtermen keine Verletzungen der Grundannahmen des Schätzansatzes feststellen. Ferner bestätigt der Jarque-Bera-Test die Normalverteilungsannahme für die Residuen und CUSUM-Tests geben keine Hinweise auf Parameter- oder Varianzinstabilität.

Prognosezeitraum	Prognoseprozedure: Fully Dynamic			
... Monate voraus	RMSE in Basispunkten	Irrtumswahrscheinlichkeit	Vorzeichentrefferquote in %	Irrtumswahrscheinlichkeit
1	24.01	39.1%	52.5	53.77%
2	34.82	45.5%	55.8	23.08%
3	41.34	48.6%	58.3	15.85%
4	47.36	37.8%	60.0	8.61%
5	52.38	30.1%	61.7	7.85%
6	56.12	31.9%	56.7	33.51%
7	58.32	38.3%	60.8	13.87%
8	59.17	40.7%	60.0	17.93%
9	59.18	41.7%	65.8	2.77%
10	59.50	40.0%	67.5	1.48%
11	60.85	45.0%	70.8	0.27%
12	62.63	43.3%	70.8	0.49%

Abbildung 11: RMSE und Vorzeichentrefferquote des erweiterten SEECM

Mit diesem neuen SEECM führen wir nun auch Out-of-Sample-Tests in der schon in Kapitel 3 beschriebenen Art durch. Die Ergebnisse dazu finden sich in Abbildung 11. Vergleicht man RMSE und Vorzeichentrefferquote mit den Ergebnissen in Abbildung 5, sieht man, dass der RMSE sich meistens leicht verringert hat, während sich die Vorzeichentrefferquoten allesamt sehr ähneln. Das heißt, dass die Hinzunahme der monetären Basis als weitere Erklärungsvariable prognostisch keine großartige Verbesserung bringt. Darüber hinaus zeigen Diebold-Marino-Tests[23], dass der RMSE des erweiterten SEECM nicht signifikant besser

[21] Beltran et. al. [16] stellen in ihrer Untersuchung auch einen signifikanten Effekt der Auslandsnachfrage auf die US-Staatsanleiherenditen fest (siehe auch Kaminska/Zinna [17]). Dieser Effekt scheint auch für das Euroland zu gelten, siehe Carvalho/Fidora [18].

[22] Kaminska/Zinna [17] kommen in ihrer Analyse zu einer ähnlichen Größenordnung, allerdings stellen sie bei ihrer Schätzung auf die Realrendite ab.

[23] Zur Berechnung der Irrtumswahrscheinlichkeit des RMSE siehe Diebold/Mariano [19].

ist als ein einfacher Random Walk (ohne Drift). Allerdings ist die Vorzeichentrefferquote des erweiterten SEECM ab einem Prognosezeitraum von mehr als neun Monaten besser als ein „Münzwurf". Ab diesem Zeitpunkt ist die Vorzeichentrefferquote signifikant größer als 50% (Irrtumswahrscheinlichkeit < 5%, auf 10%-Niveau auch für Monate 4 und 5).[24] Zusammenfassend deuten die Out-of-Sample-Tests darauf hin, dass die Prognose der US-Staatsanleihenzinsen auf Basis von Makro-Variablen seit der unorthodoxen Geldpolitik der FED schwieriger geworden ist.[25]

5. Zusammenfassung

Seit der Finanzkrise verfolgt die FED einen sehr expansiven geldpolitischen Kurs. Die Zinsen wurden auf historische Niedrigstniveaus gesenkt, darüber hinaus griff die FED auf verschiedene Maßnahmen quantitativer Lockerungen zurück. Trotz dieser Ereignisse zeigen die ökonometrischen Analysen in diesem Papier, dass die klassischen Einflussfaktoren (Notenbankzinsen, Inflationserwartungen, Auslandsnachfrage nach US-Staatsanleihen) weiterhin gute Erklärungskraft haben. Allerdings sind diese Zusammenhänge in jüngster Zeit nicht mehr so stabil. Darüber hinaus kann man feststellen, dass das Anleiheaufkaufprogramm der amerikanischen Notenbank wohl einen messbaren Einfluss auf den zehnjährigen US-Staatsanleihenzins hat.

Die Aufgabe der zukünftigen Forschung ist es, zu prüfen, ob es sich um vorübergehende Instabilitäten handelt oder nach dem Ende der expansiven unorthodoxen Geldpolitik der FED die ursprünglichen, sehr stabilen fundamentalen Zusammenhänge zwischen Zinsen und seinen Fundamentalfaktoren wieder gültig sind.

Ob QE ein wirkungsvolles geldpolitisches Instrumentarium ist, kann nicht letztendlich und generell beantwortet werden. Im Ansatz wurde nur untersucht, ob die FED-Politik auf die Zinsen wirkt. Die Antwort hierfür kann wohl bejaht werden. Ob die FED damit allerdings die Kreditnachfrage, die Güternachfrage und die Konjunktur stimulieren konnte, bleibt ebenfalls weiterer Forschung vorbehalten.

6. Disclaimer

Alle Meinungsaussagen oder Einschätzungen in diesem Dokument geben die Einschätzung des Verfassers bzw. der Verfasser wieder. Die hierin zum Ausdruck gebrachten Meinungen spiegeln nicht zwangsläufig die Meinungen der UniCredit Bank AG wider. Verantwortlich für den Inhalt ist allein der jeweilige Autor.

7. Abkürzungen

BCS-Modell:
Zinsmodell von Bandholz et. al. (2009)

BIZ:
Bank für Internationalen Zahlungsausgleich

CORE:
US Core Inflation (Headline Consumer Price Index ohne Nahrungsmittel und Energie)

EZB:
Europäische Zentralbank

FED:
Federal Reserve System (US-Notenbank)

HOLDING:
Anteil der vom Ausland gehaltenen US-Staatsanleihen gebildet aus dem Verhältnis von „Rest of the World Treasury Securities" und „Federal Government Debt Securities"

ISM:
Index des Verarbeitenden Gewerbes vom Institute for Supply Management

IWF:
Internationaler Währungsfonds

K_MB_GDP:
Das Verhältnis zwischen der monetären Basis der FED und dem nominalen US-BIP

KZINS:
US-Geldmarktzins für Dreimonatsgeld

LZINS:
Zins für zehnjährige US-Staatsanleihen

[24] Zur Berechnung der Irrtumswahrscheinlichkeit der Vorzeichentrefferquote siehe Cheung et. al. [20].
[25] Bauer/Hamilton [21, 22] stellen grundsätzlich in Frage, ob eine Prognose von Zinsen auf Basis von Makro-Variablen überhaupt möglich ist.

OECD:
Organization of Economic Cooperation and Development

QE:
Quantitative Easing

RMSE:
Root Mean Squared Error

SEECM:
Einzelgleichungsfehlerkorrekturmodell

VECM:
Vektorfehlerkorrekturmodell

8. Literatur

[1] Bandholz, Harm; Clostermann, Jörg; Seitz, Franz (2009): "Explaining the US Bond Yield Conundrum". In: *Applied Financial Economics* 19 (7), S. 539–550.

[2] Belke, Ansgar; Gros, Daniel; Osowski, Thomas (2016): "Did Quantitative Easing Affect Interest Rates Outside the US? – New Evidence Based on Interest Rate Differentials. Federal Reserve Bank of San Francisco (FRBSF Economic Letters 2016-20). [Online verfügbar]. https://www.ceps.eu/system/files/WD416%20Did%20QE%20affect%20interest%20rates_0.pdf. (zuletzt geprüft am 29.09.2016).

[3] Fawley, Brett W.; Neely, Christopher J. (2013): "Four Stories of Quantitative Easing." In: *Federal Reserve Bank of St. Louis Review* 95 (1), S. 51–88.

[4] Greenspan, A. (2005): Testimony of Chairman Alan Greenspan. Federal Reserve Board's semiannual Monetary Policy Report to the Congress. Before the Committee on Banking, Housing, and Urban Affairs of the U.S. Senate. Washington, DC, 16.02.2005.

[5] Bernanke, Ben S. (March 20th 2006): Remarks by Ben S. Bernanke (Chairman of the Board of Governors of the Federal Reserve System) before the Economic Club of New York. New York, NY, 20.03.2006.

[6] Obstfeld, Maurice; Tesar, Linda (2015): The Decline in Long-Term Interest Rates, the WHITE HOUSE President Barack Obama. The White House/President Barack Obama. July 14, 2015 at 5:35 PM ET. [Online verfügbar]. https://www.whitehouse.gov/blog/2015/07/14/decline-long-term-interest-rates. (zuletzt geprüft am 18.08.2016).

[7] The Executive Office of the President of the United States (2015): Long Term Interest-Rate: A Survey. Council of Economic Advisors. [Online verfügbar]. https://www.whitehouse.gov/sites/default/files/docs/interest_rate_report_final.pdf,. (zuletzt geprüft am 18.08.2016).

[8] Auer, Benjamin; Rottmann, Horst (2015): *Statistik und Ökonometrie für Wirtschaftswissenschaftler. Eine anwendungsorientierte Einführung.* 3., überarb. u. aktualisierte Aufl. 2015. Wiesbaden: Springer Gabler (SpringerLink : Bücher).

[9] Issing, Otmar (2011): *Einführung in die Geldtheorie.* 15., wesentlich überarbeitete Auflage. München: Vahlen (Vahlens Kurzlehrbücher).

[10] Board of Governors of the Federal Reserve System (2016): Z.1 Financial Accounts of the United States. Flow of Funds, Balance Sheets, and Integrated Macroeconomic Accounts. Second Quarter 2016. [Online verfügbar]. https://www.federalreserve.gov/releases/z1/current/z1.pdf. (zuletzt geprüft am 18.08.2016.

[11] Wu, Tao (2014): Unconventional Monetary Policy and Long-Term Interest Rates. IMF Working Paper WP/14/189. International Monetary Fund/Institute for Capacity Development. [Online verfügbar]. https://www.imf.org/external/pubs/ft/wp/2014/wp14189.pdf. (zuletzt geprüft am 18.08.2016).

[12] Banerjee, Anindya; Dolado, Juan; Mestre, Ricardo (1998): "Error-Correction Mechanism Tests for Cointegration in a Single-Equation Framework." In: *Journal of Time Series Analysis* 19 (3), S. 267–283.

[13] Engle, Robert F.; Hendry, David F.; Richard, Jean-François (1983): "Exogeneity." In: *Econometrica* (Hg.: The Econometric Society) 51 (2), S. 277–304.

[14] Johansen, Søren (1992): "Cointegration in Partial Systems and the Efficiency of Single-Equation Analysis." In: *Journal of Econometrics* 52 (3), S. 389–402.

[15] Derksen, Shelley; Keselmann, H. J. (1992): "Backward, Forward and Stepwise Automated Subset Selection Algorithms: Frequency of Obtaining Authentic and Noise Variables." In: *British Journal of Mathematical and Statistical Psychology* 45 (2), S. 265–282.

[16] Beltran, Daniel O.; Kretchmer, Maxwell; Marquez, Jaime; Thomas, Charles P. (2013): "Foreign Holdings of U.S. Treasuries and U.S. Treasury Yields." In: *Journal of International Money and Finance* 32 (February), S. 1120–1143.

[17] Kaminska, Iryna; Zinna, Gabriele (2014): Official Demand for U.S. Debt: Implications for U.S. Real Interest Rates. IMF Working Paper WP/14/66. International Monetary Fund. [Online verfügbar]. https://www.imf.org/external/pubs/ft/wp/2014/wp1466.pdf. (zuletzt geprüft am 18.08.2016).

[18] Carvalho, Daniel; Fidora, Michael (2015): Capital Inflows and Euro Area Long-Term Interest Rates. June 2015. European Central Bank (ECB Working Paper 1798).[Online verfügbar]. http://www.ecb.europa.eu/pub/pdf/scpwps/ecbwp1798.en.pdf. (zuletzt geprüft am 29.09.2016).

[19] Diebold, Francis X.; Mariano, Roberto S. (1995): "Comparing Predictive Accuracy." In: J*ournal of Business & Economic Statistics* 13, S. 253–263.

[20] Cheung, Yin-Wong; Chinn, Menzie D.; Pascual, Antonio Garcia (2005): "Empirical Exchange Rate

Models of the Nineties: Are Any Fit to Survive?" In: *Journal of International Money and Finance* 24, S. 1150–1175.

[21] Bauer, Michael D.; Hamilton, James D. (2015): Robust Bond Risk Premia. Federal Reserve Bank of San Francisco (Working Paper Series, Working Paper 2015-15). [Online verfügbar]. http://www.frbsf.org/economic-research/files/wp2015-15.pdf. (zuletzt geprüft am 18.08.2016).

[22] Bauer, Michael D.; Hamilton, James D. (2016): Do Macro Variables Help Forecast Interest Rates? Federal Reserve Bank of San Francisco (FRBSF Economic Letter 2016-20). [Online verfügbar]. http://www.frbsf.org/economic-research/publications/economic-letter/2016/june/do-macroeconomic-variables-help-forecast-interest-rates/. (zuletzt geprüft am 29.09.2016)

Anhang

Anhang 1: Augmented-Dickey-Fuller-Tests

Zeitraum: Jan 1986 - Dez. 2015

Variable	Spezifikation[1)]	Lag[2)]	t-Wert[3)]
IN(ISM)	C	3	-6.06 **
LZINS	N	0	-1.70
D(LZINS)	N	0	-17.75 **
KZINS	N	2	-1.66
D(KZINS)	N	1	-9.87 **
CORE	N	12	-1.58
D(CORE)	N	11	-7.05 **
HOLDING	T	16	-1.89
D(HOLDING)	N	15	-2.26 *
K_MB_GDP	T	2	-0.83
D(K_MB_GDP)	C	0	-11.04 **

1) T = mit Konstante und Trend; C = mit Konst., ohne Trend; N = keine Konst., kein Trend.
2) Lag-Selektion basierend auf dem Schwarz Informationskriterium. Maximale Laglänge = 16.
3) **/* = 1/5% Signifikanzniveau.

Anhang 2: Kwiatkowski-Phillips-Schmidt-Shin-Tests

Zeitraum: Jan 1986 - Dez. 2015

Variable	Spezifikation[1)]	Bandwith[2)]	Teststatistik[3)]
IN(ISM)	C	14	0.05
LZINS	C	15	2.20 **
LZINS	T	14	0.04
D(LZINS)	C	7	0.02
KZINS	C	15	-1.72 **
KZINS	T	15	0.06
D(KZINS)	C	11	0.04
CORE	C	15	1.78 **
CORE	T	15	0.27 **
D(CORE)	C	8	0.08
HOLDING	T	15	0.19 *
D(HOLDING)	C	5	0.22
K_MB_GDP	T	15	0.46 **
D(K_MB_GDP)	T	10	0.08
D(K_MB_GDP)	C	10	0.48 *

1) T = mit Konstante und Trend; C = mit Konst., ohne Trend.
2) Automatische Newey-West Korrektur auf Basis des Bartlett-Kernel.
3) Kwiatkowski-Phillips-Schmidt-Shin Teststatistik; **/* = 1/5% Signifikanzniveau.

Anhang 3: VECM mit Restriktionen auf schwache Exogenität (t-Werte in Klammern)

Kointegrationsgleichung	
$LZINS_{-1}$	1.00000
$KZINS_{-1}$	-0.26228
	[-3.85817]
$HOLDING_{-1}$	7.23917
	[4.27037]
$CORE_{-1}$	-0.48146
	[-3.06344]
$K\_MB\_GDP_{-1}$	6.20925
	[2.35905]
Konstante	-5.66602

Kurzfristgleichung	ΔLZINS	ΔKZINS	ΔHOLDING	ΔCORE	ΔK_MB_GDP
Fehlerkorrekturen	-0.16136	0.00000	0.00000	0.00000	0.00000
	[-5.78053]	[NA]	[NA]	[NA]	[NA]
Δ $LZINS_{-1}$	0.10718	0.06176	0.00052	-0.02280	-0.00035
	[1.95969]	[1.39209]	[1.01710]	[-0.87248]	[-0.69255]
Δ $KZINS_{-1}$	-0.13454	0.02003	-0.00157	0.02866	-0.00057
	[-1.89153]	[0.34711]	[-2.33950]	[0.84334]	[-0.87340]
Δ $HOLDING_{-1}$	-9.14064	-6.58571	0.61065	-0.80227	0.05183
	[-2.03180]	[-1.80459]	[14.4077]	[-0.37322]	[1.26118]
Δ $CORE_{-1}$	0.14550	-0.04949	-0.00081	0.11365	-0.00079
	[1.31204]	[-0.55009]	[-0.77012]	[2.14486]	[-0.77544]
Δ $K\_MB\_GDP_{-1}$	14.98849	5.15237	-0.00329	-0.13973	0.42425
	[2.75153]	[1.16599]	[-0.06400]	[-0.05368]	[8.52616]
Konstante	-2.84499	-2.92730	-0.01245	-0.33446	0.01178
	[-4.00017]	[-5.07384]	[-1.85798]	[-0.98420]	[1.81341]
LN(ISM)	2.27908	1.01163	-0.00703	-0.03552	-0.00602
	[5.60266]	[3.06569]	[-1.83524]	[-0.18276]	[-1.61950]
$Ln(ISM_{-1})$	-1.56301	-0.27398	0.01025	0.11907	0.00308
	[-3.91873]	[-0.84679]	[2.72755]	[0.62477]	[0.84579]
R^2	0.15978	0.14291	0.41177	0.03741	0.24740
Adj. R^2	0.14063	0.12337	0.39836	0.01547	0.23025
Standardfehler	0.26517	0.21510	0.00250	0.12670	0.00242
F-Wert	8.34367	7.31550	30.71282	1.70510	14.42294

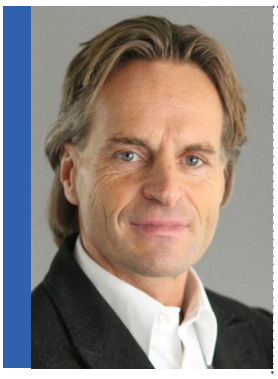
Dr. Harm Bandholz

Dr. Harm Bandholz ist US-Chefvolkswirt der UniCredit Bank in New York. In seiner Tätigkeit analysiert er sowohl konjunkturelle Entwicklungen als auch geld- und fiskalpolitische Tendenzen, insbesondere solche mit Relevanz für die Finanzmärkte. Vor seinem Wechsel zur UniCredit arbeitete Herr Bandholz als wissenschaftlicher Mitarbeiter am Ifo-Institut in München sowie am Lehrstuhl für angewandte Makroökonomie der Universität Hamburg. Er ist Mitglied des Economic Club of New York, des Forecasters Club of New York, des American Council on Germany, der New York Society of Security Analysts und des CFA Institute.

Harm Bandholz is the Chief US Economist for UniCredit. In his role, he analyzes the macroeconomy as well as monetary and fiscal policy developments, in particular those with relevance for financial markets. Before joining the bank, Dr. Bandholz worked as a Research Assistant in the Department of Business Cycle Analyses & Financial Markets of the Ifo Institute for Economic Research in Munich, and as a Teaching and Research Assistant at the Department of Economics at Hamburg University. He is a member of the Economic Club of New York, the Forecasters Club of New York, the American Council on Germany, the New York Society of Security Analysts, and the CFA Institute.

Kontakt / *Contact*

harm.bandholz@unicredit.eu



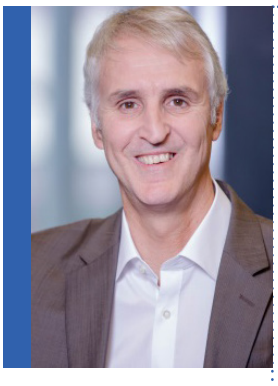
Prof. Dr. Jörg Clostermann

Prof. Dr. Jörg Clostermann lehrt Volkwirtschaftslehre und Quantitative Methoden an der Technischen Hochschule Ingolstadt. Vor seiner Berufung arbeitete er in der Hauptabteilung Volkwirtschaft der Deutschen Bundesbank in Frankfurt, Abteilung Außenwirtschaft. Prof. Clostermann berät seit vielen Jahren Banken und Unternehmen auf dem Gebiet Indikatoren und Prognosen. Er ist darüber hinaus Mitglied des Aktionskreises Stabiles Geld.

Professor Joerg Clostermann teaches Economics and Statistics at Ingolstadt Technical University of Applied Sciences. Before his professorship, he worked at the Deutsche Bundesbank in the division Economics and the department International Economics. For many years now, Professor Clostermann has been acting as a consultant for banks and enterprises in the area of economic indicators and economic forecasts. Furthermore, he is a member of the Aktionskreis: Stabiles Geld.

Kontakt / *Contact*

✉ **joerg.clostermann@thi.de**



Prof. Dr. Franz Seitz

Prof. Dr. Franz Seitz lehrt Volkswirtschaftslehre, insbesondere Geldpolitik und Finanzmärkte, an der Ostbayerischen Technischen Hochschule in Weiden. Zugleich ist er federführendes Mitglied im Aktionskreis Stabiles Geld. Vor seiner Berufung als Professor arbeitete er in der Hauptabteilung Volkswirtschaft der Deutschen Bundesbank, Abteilung Geld, Kredit und Kapitalmarkt, in Frankfurt am Main. Prof. Seitz ist Autor zahlreicher Artikel in nationalen und internationalen Fachzeitschriften. Seit etlichen Jahren fungiert Prof. Seitz in mehreren Projekten als Berater der Europäischen Zentralbank und der Deutschen Bundesbank.

Professor Franz Seitz teaches Economics with a special focus on Monetary Policy and Financial Markets at Weiden Technical University of Applied Sciences. He is also one of the leading members of the Aktionskreis Stabiles Geld. Before his professorship, he worked at the Deutsche Bundesbank in the division Money, Credit and Capital Markets. Professor Seitz is author of numerous articles in national and international journals. For many years now, Professor Seitz has been acting as a consultant in different projects for the Deutsche Bundesbank and the European Central Bank.

Kontakt / *Contact*

✉ **f.seitz@oth-aw.de**

BAVARIAN JOURNAL OF APPLIED SCIENCES®

Wunsch und Wirklichkeit verbindlicher Pflegeplanung - Motive und Wirkungen des nordrheinwestfälischen Sonderweges

Hanjo Allinger
Technische Hochschule Deggendorf

ABSTRACT

In Nordrhein-Westfalen können Kreisverwaltungen die Ergebnisse ihrer Pflegebedarfsplanung für verbindlich erklären. Geht man behördlich davon aus, dass mehr Pflegebetten angeboten als benötigt werden, kann für Bewohner von neu gebauten Einrichtungen das Pflegewohngeld gestrichen werden. Bei einer Unterversorgung werden dagegen keine zusätzlichen Anreize für weitere Investitionen gesetzt. Gezielt wird damit ausschließlich auf eine Begrenzung des Angebots stationärer Pflegebetten. Liegt ein derartiger Markteingriff im Interesse der Allgemeinheit? Die ökonomische Literatur kennt verschiedene Konstellationen, die regelmäßig zu einem Versagen von freien Märkten führen. Durch staatliche Regulierung könnte in diesen Fällen oft die gesamtgesellschaftliche Wohlfahrt gestärkt werden. Derartige Begründungen für den Markteingriff sind im Bereich der stationären Pflege jedoch nicht ersichtlich. Tatsächlich erscheint es plausibler, dass Kreise mit verbindlicher Pflegeplanung von rein finanziellen Erwägungen getrieben werden. Aber auch diese beruhen entweder auf einer Fehleinschätzung der finanziellen Auswirkungen oder auf rechtlich unhaltbaren Zielen wie dem klar wettbewerbswidrigen Schutz öffentlicher Betreiber vor privater Konkurrenz.
Tatsächlich spricht viel dafür, dass das Landespflegegesetz mit der selektiven Verweigerung des Pflegewohngeldes die nach Art. 12 I GG garantierte Berufsfreiheit verfassungswidrig einschränkt.

In North Rhine-Westphalia, district administrations can declare the results of their care requirements planning to be binding. If local authorities assume that more places in nursing homes are offered than needed, care housing benefits for residents in newly built facilities can be cancelled. In the case of an under-supply, however, no additional incentives are set for further investments. The sole aim is therefore to limit the supply of nursing home beds. Is such a market intervention in the public interest? The economic literature knows various constellations which regularly lead to failures of free markets. In these cases, state intervention and regulation can often increase the overall welfare.

However, such justifications for market interventions are not apparent in the area of inpatient care. In fact, it seems more plausible that district administrations with mandatory care planning are driven by purely financial considerations. However, these are either based on a misperception of financial implications or on forbidden targets such as the anti-competitive protection of public operators against private competition.
In fact, there is much evidence that the Landespflegegesetz, with the selective refusal of care housing benefits for residents of nursing homes, unconstitutionally limits the professional freedom which is protected according to Art. 12 I GG.

KEYWORDS

Stationäre Pflege, verbindliche Pflegeplanung, Pflegewohngeld, Investitionskostenförderung, Überangebot

Inpatient care, mandatory care planning, care housing benefit, public funding of investment costs, oversupply

1. Hintergrund

In Nordrhein-Westfalen wurde 2014 das Landespflegegesetz durch das Alten- und Pflegegesetz (APG NRW) ersetzt. Dieses Gesetz verpflichtet die Kreise und kreisfreien Städte zur Planung des stationären Pflegebedarfs[1]. Anders als in anderen Bundesländern können Kreise und kreisfreie Städte in Nordrhein-Westfalen diese Pflegeplanung per Beschluss für verbindlich erklären. Liegt die bereits angebotene Anzahl von Pflegebetten in einer Gemeinde über den – wie auch immer ermittelten – Planzahlen der Verwaltung, wird den Bewohnern neu errichteter Pflegeheime das Pflegewohngeld versagt.[2]

Das Pflegewohngeld wird in NRW Pflegebedürftigen gewährt, die sozialhilfeberechtigt sind oder deren Einkommen und Vermögen nicht ausreicht, die Investitionskosten der Pflegeeinrichtung selbst in voller Höhe zu tragen.[3] Die Höhe des Pflegewohngeldes entspricht maximal den betriebsnotwendigen Investitionsaufwendungen[4] abzüglich des anrechenbaren Einkommens des Pflegebedürftigen.[5] Im Unterschied zur Sozialhilfe gilt hier mit 10.000 Euro Barvermögen ein großzügigeres Schonvermögen sowie geringfügig gelockerte Anrechnungsvorschriften beim eigenen Einkommen.[6] Zu einer fehlerhaften Einordnung des Pflegegeldes könnte der Umstand führen, dass das Pflegewohngeld nicht an den leistungsberechtigten Bedürftigen ausgezahlt wird, sondern an die Pflegeeinrichtung.[7] Zu allem Überfluss ist der entsprechende Paragraph 14 des Alten- und Pflegegesetzes auch noch mit „Förderung vollstationärer Dauerpflegeeinrichtungen (Pflegewohngeld)“ überschrieben. Beides ändert jedoch nichts an der Tatsache, dass die finanzielle Entlastungswirkung des Pflegegeldes eindeutig bei dem Pflegebedürftigen und nicht bei der Pflegeeinrichtung liegt: Wird eine Pflegeeinrichtung von der Gewährung des Pflegewohngeldes ausgeschlossen, erhöht sich die vom Pflegebedürftigen bzw. dem Sozialhilfeträger zu tragende Rechnungsbetrag um exakt den verwehrten Betrag.

Über den Umweg des Pflegebedürftigen und damit auch auf dessen Rücken wird versucht, die Wettbewerbsposition von Pflegeeinrichtungen, die über den Planbedarf hinaus Betten anbieten, zu verschlechtern. Durch den Wegfall der finanziellen Zuwendungen an den Pflegebedürftigen soll die Attraktivität einer Pflegeeinrichtung so weit gesenkt werden,

[1] Vgl. § 7 APG NRW
[2] Vgl. § 11 Abs. 7 APG NRW
[3] Vgl. § 14 Abs. 1 APG NRW
[4] Vgl. § 14 APG DVO NRW
[5] Vgl. § 14 Abs. 3 APG NRW
[6] Vgl. § 14 Abs. 3 APG NRW
[7] Vgl. § 16 Abs. 4 DVO APG NRW

dass die Nachfrage nach den nicht geförderten Pflegebetten erwartbar zurückgeht und der Neubau einer Pflegeeinrichtung bereits im Vorfeld für den Betreiber unattraktiv erscheint und gar nicht erst gebaut wird.

Drei Fragen stellen sich in diesem Zusammenhang, denen dieser Beitrag nachgehen möchte.

1) Gibt es wohlfahrtsökonomische oder volkswirtschaftliche Gründe für einen staatlichen Markteingriff zur Begrenzung des Angebots an Pflegeplätzen?

2) Welche finanziellen Erwägungen könnten hinter einem solchen Staatseingriff stehen?

3) Kann mit der selektiven Gewährung von Pflegewohngeld überhaupt das Angebot an Pflegeplätzen beeinflusst werden und ist dies rechtlich zulässig?

2. Volkswirtschaftliche Motive für eine Begrenzung des Angebots an Pflegeplätzen sind nicht erkenntlich

Volkswirtschaftliche Gründe für einen Staatseingriff in den Pflegemarkt würden dann vorliegen, wenn durch die Bekämpfung eines bestehenden Marktversagens gesamtgesellschaftliche Wohlfahrtsgewinne erzielt werden können.

2.1 Angebotsinduzierte Nachfrage im Bereich der stationären Pflege nicht ersichtlich

Im Bereich der niedergelassenen Ärzte wird häufig eine angebotsinduzierte Nachfrage als Grund für ein Versagen der Märkte und die Notwendigkeit staatlicher Eingriffe in die soziale Marktwirtschaft durch eine Begrenzung der Kassenzulassungen angeführt. Auch wird in diesem Bereich gelegentlich darauf verwiesen, dass die leichtere Verfügbarkeit von Leistungen auch die Anzahl der nachgefragten Leistungen erhöhen könne. Beide Argumente greifen im Bereich der stationären Pflege nicht. Das Argument der angebotsinduzierten Nachfrage speist sich im ambulanten Sektor letztlich aus Informationsasymmetrien: Der Patient kann nicht beurteilen, welche diagnostischen Verfahren medizinisch für seine Behandlung notwendig sind, und dem Arzt steht eine Vielzahl möglicher Verfahren unterschiedlicher Güte und unterschiedlicher Kosten zur Verfügung, aus denen er wählen kann. In der stationären Pflege gibt es aber kaum eine Möglichkeit für den Leistungsanbieter, den Umfang der gewährten Leistungen zu Lasten der Krankenkassen auszuweiten. Die Einstufung in Pflegestufen erfolgt unabhängig durch den medizinischen Dienst der Krankenkassen und auch die medizinische Notwendigkeit der stationären Pflege wird im Regelfall zunächst durch eine Pflegekasse festgestellt. Ist dies nicht der Fall, werden nach §43 Abs. 4 SGB XI nur die deutlich geringeren Sätze für häusliche Pflege als Zuschuss für die Heimunterbringung gewährt. Wegen dieser finanziellen Schlechterstellung ist eine „Selbsteinweisung“ für Pflegebedürftige wenig wahrscheinlich.

2.2 Moral Hazard bei der Nachfrage nach Pflegebetten unplausibel

Die zweite Befürchtung, höhere Kapazitäten könnten die Nachfrage erhöhen, greift gerade nicht bei einer Überversorgung, sondern lediglich bei einer Unterversorgung mit Heimplätzen. Bei zu wenig Plätzen in stationärer Dauerpflege wird ersatzweise auf Leistungen der häuslichen Pflege zurückgegriffen, die dann bei einer Aufstockung der Kapazitäten zu einer Mehrnachfrage bei neu verfügbaren Heimplätzen führen. Denkbar ist auch, dass der Umzug vom eigenen Hausstand in eine Pflegeeinrichtung durch einen modernen Neubau im Heimatort erleichtert wird. Aber auch hier ist die Pflegebedürftigkeit exogen gegeben und eine Erleichterung der für Pflegebedürftige und deren Angehörige schwierigen Situation ist eine Pflicht der Solidargemeinschaft.

2.3 Marktzutrittsbeschränkungen führen zu Investitionsstau

Welcher Wohlfahrtsverlust sollte entstehen, wenn das Angebot an Pflegeplätzen größer als die Nachfrage ist? Dass Überkapazitäten die Qualität der Versorgung nachträglich beeinträchtigen, ist nicht zu befürchten – schließlich sind die Mindestversorgungsstandards in der Pflege unabhängig von der Auslastung einer Einrichtung allesamt per Gesetz vorgegeben. Eine Einhaltung dieser Vorgaben kann durch Kontrollen der Aufsichtsbehörden im verfassungsrechtlichen Sinne verhältnismäßiger überprüft werden.

Zu Recht fühlt sich bei einem Überangebot von Apotheken keine Kreisverwaltung (mehr) für dessen Begrenzung zuständig – ganz klar

wird dies seit dem wegweisenden „Apotheken-Urteil“ des Bundesverfassungsgerichts [1] als Aufgabe des Marktes gesehen. Das Bundesverfassungsgericht konnte damals keinen Beleg dafür finden, dass selbst ein ruinöser Wettbewerb unter Apotheken die Volksgesundheit gefährden würde. „Nur die Abwehr nachweisbarer oder höchstwahrscheinlicher, schwerer Gefahren für ein überragend wichtiges Gemeinschaftsgut“, wie dies auch die Volksgesundheit sei, könne eine derart massive Einschränkung der in Art. 12 I GG garantierten Berufsfreiheit durch die Einführung objektiver Zulassungsvoraussetzungen rechtfertigen [2]. Aus ökonomischer Sicht macht erst das Überangebot Wettbewerb zwischen Anbietern möglich. Und nur durch den Wettbewerb besteht für die Anbieter ein Anreiz, ihre Leistungen zu verbessern. Würden tatsächlich durch die Streichung des Pflegewohngeldes bei „Planübererfüllung“ und die damit verbundenen Wettbewerbsnachteile Neuanbieter abgeschreckt, würden nicht nur keine Neubauten mehr entstehen, sondern mangels Wettbewerbsdrucks auch Investitionen in bestehende Einrichtungen unterbleiben. In der Folge überaltert die Bausubstanz, Pflegestandards sinken. Als 2003 die Pflegeplanung in NRW unter anderem wegen der Bedenken des Bundessozialgerichts gegen die Bindung der Investitionskostenförderung an eine Bedarfsfeststellung abgeschafft wurde[8], berechneten die Landschaftsverbände einen Investitionsrückstand von 4,7 Mrd. Euro [3].
Nicht die Beschränkung des Wettbewerbs, sondern seine Stärkung wäre im Interesse des Gemeinwohls. Aus volkswirtschaftlicher Perspektive ist schwer nachzuvollziehen, wie es im Interesse der Allgemeinheit liegen könnte, neue Pflegeheime zu verhindern.

3. Finanzielle Motive erscheinen plausibler, sind jedoch auch nicht belastbar

Anders als bei den volkswirtschaftlichen Motiven geht es bei primär betriebswirtschaftlichen oder finanziellen Motiven darum, im Rahmen von marktwirtschaftlichen Verteilungskämpfen individuell besser abzuschneiden. Salopp formuliert steht hier nicht die Größe des zur Verfügung stehenden Kuchens, sondern die Größe des eigenen Kuchenstücks im Mittelpunkt des Interesses.

3.1 Schutz kommunaler Einrichtungen vor Wettbewerbern

Ein betriebswirtschaftliches Motiv scheint bei einigen nordrheinwestfälischen Kreisverwaltungen, die für eine verbindliche Pflegeplanung optieren, nicht unplausibel: der Schutz bestehender und häufig in die Tage gekommener kommunaler Pflegeeinrichtungen vor privaten Mitbewerbern. Dieses Motiv der Wettbewerbsbeschränkung zugunsten öffentlicher Pflegeeinrichtungen ist klar rechtswidrig und wird bei erfolgreicher Beweisführung durch die zuständigen Gerichte sanktioniert werden.

3.2 Vermeidung des Zuzugs von Sozialhilfeempfängern

Mit Hilfe der verbindlichen Pflegeplanung sollen durch die wettbewerbliche Schlechterstellung neuer Anbieter Überkapazitäten vermieden werden. Dadurch – so hofft man in einigen Kreisen – könne verhindert werden, dass durch freie Pflegekapazitäten im eigenen Kreis Pflegebedürftige aus stärker ausgelasteten Nachbarkreisen angelockt werden und für die stationäre Pflege ihren Wohnort in den Kreis verlegen. Befürchtet wird, dass diese „Pflege-Neubürger“ zu Mehrkosten im Sozialhaushalt führen, wenn sie eines Tages im eigenen statt im Heimatkreis Sozialhilfe beantragen würden.
So verquast sich dieses Argument auf den ersten Blick auch anhört, tatsächlich bestätigten dem Autor mehrere Kreisverwaltungen in Nordrhein-Westfalen, dass genau dies ihre größte Sorge und das Hauptargument dafür sei, die Pflegeplanung für verbindlich zu erklären.
De facto baut diese Argumentation jedoch weitgehend auf einer unzulänglichen Rechtskenntnis auf. Mit dem Einzug in ein Pflegeheim eines anderen Kreises wird nämlich gerade nicht, wie in der Argumentation unterstellt, automatisch auch der Sozialhilfeträger des neuen Wohnorts zuständig. Zieht ein Pflegebedürftiger aus einem anderen Landkreis in eine stationäre Pflegeeinrichtung, bleibt nach § 98 Abs. 2 SGB XII weiterhin das auswärtige Sozialamt, in dem der Hilfebedürftige in den zwei Monaten vor der Aufnahme seinen gewöhnlichen Aufenthalt gehabt hat, leistungsverpflichtet. Das gilt sowohl für Bezieher von Sozialhilfeleistungen als auch für den Fall, dass sich der Sozialhilfebedarf

[8] Vgl. Ausführungen zur Gesetzesbegründung unter A 3.in [4].

erst Jahre nach dem Umzug in das Pflegeheim manifestiert. Die Leistungsverpflichtung geht somit nur dann über, wenn der gewöhnliche Aufenthaltsort bereits mindestens zwei Monate vor Heimeintritt in den Kreis verlagert wurde – etwa in eine ambulant betreute Wohngemeinschaft. Diese Fälle sind jedoch erstens zahlenmäßig zu vernachlässigen und zweitens durch die Bekämpfung freier Kapazitäten in der stationären Pflege nicht zu verhindern.

3.3 Hoffnung auf Einsparungen

Kreisverwaltungen könnten sich womöglich Hoffnung auf direkte Einsparungen durch den Wegfall der Pflegewohngeldzahlungen machen. Tatsächlich ist jedoch auch diese Hoffnung weitgehend unbegründet: Da die betriebsnotwendigen Investitionskosten einer stationären Einrichtung beim Wegfall des Pflegewohngeldes den Pflegebedürftigen direkt in Rechnung gestellt werden dürfen[9], führt dies bei Sozialhilfeempfängern über den Umweg des gestiegenen Hilfsbedarf des Pflegebedürftigen im selben Umfang zu Mehrleistungen der Sozialhilfe. Eine Ausnahme sieht § 75 Abs. 5 für gesondert berechnete Investitionskosten nach § 82 Abs. 4 vor, wenn keine Vereinbarung zur Übernahme zwischen dem Träger der Sozialhilfe und der Pflegeeinrichtung geschlossen ist.

Eine juristische Hintertür für den Sozialhilfeträger wird nach übereinstimmender Auffassung einschlägiger Kommentare damit jedoch nicht geöffnet, denn zum einen hat der Einrichtungsträger einen Anspruch auf Prüfung des Abschlusses einer Vereinbarung gegenüber dem Leistungsträger [5]. Zum anderen „rechtfertigt der Umstand, dass ein Pflegeheim nach dem Landesrecht nicht gefördert wird, nicht die Entscheidung des Trägers der Sozialhilfe, den Abschluss einer Vereinbarung mit dem Heimträger nach § 75 Abs. 5 S. 3 zu verweigern [...]. Ebenso kann der Abschluss einer Vereinbarung [...] nicht mit der Erwägung abgelehnt werden, im Zuständigkeitsbereich des Trägers der Sozialhilfe seien bereits genügend geförderte Pflegeplätze oder sogar ein Überhang vorhanden (OVG Lüneburg, Urt. v. 14.3.2001 – 4 L 2155/00)“ [6]. Eine derartige Begründung würde gegen die in Art. 12 GG garantierte Berufsfreiheit verstoßen [7].

Daraus folgt, dass es zu Einsparungen für die öffentliche Hand nur durch die sehr geringen Unterschiede in den Vorschriften zur Einbringung eigenen Einkommens und Vermögens bei Pflegewohngeld und Sozialhilfe kommen kann. Während beim Pflegewohngeld Barvermögen bis 10.000 Euro nicht eingebracht werden muss, ist der Freibetrag in der Sozialhilfe für Menschen über 60 Jahre nur 2.600 € hoch.[10] Für das monatliche Einkommen gilt darüber hinaus beim Pflegewohngeld ein zusätzlicher Selbstbehalt von 50 Euro. Die des Weiteren angeführten Absetzbeträge für Unterkunft und Verpflegung sowie die nicht von den Pflegekassen abgedeckten Pflegkosten müssen von der Sozialhilfe gleichermaßen übernommen werden. Die Differenz zwischen beiden Beträgen für das Barvermögen und die 50 Euro monatlich beziffern den theoretisch maximal zusätzlichen Konsumspielraum von Pflegewohngeldempfängern gegenüber Sozialhilfeempfängern und damit die maximalen Einsparungen für die öffentliche Hand. Die Mehrzahl der Pflegewohngeldempfänger kann diesen Spielraum mangels Vermögens jedoch bei weitem nicht ausschöpfen.

Bei einkommenslosen pflegebedürftigen Sozialhilfeempfängern könnte die öffentliche Hand durch den Wegfall des Pflegewohngeldes also nicht einen Euro einsparen, da diese weder über das anrechnungsfreie Barvermögen in Höhe von 10.000 Euro verfügen, noch ein laufendes Einkommen haben, bei dem die Unterschiede in der monatlichen Anrechnung zum Tragen kommen. Gleiches gilt für die große Gruppe der vermögenderen Pflegebedürftigen, die ohnehin weder sozialhilfe- noch pflegewohngeldberechtigt sind.

4. Selektiv verweigertes Pflegewohngeld ist zur Pflegebetten-Steuerung ungeeignet und rechtlich unzulässig

Wenn weder bei einkommenslosen Sozialhilfeempfängern noch bei Pflegebedürftigen ohne Hilfebedarf durch den Wegfall des Pflegewohngeldes nennenswerte Einsparungen für die öffentliche Hand erzielt werden können, bedeutet dies umgekehrt aber auch, dass den Bezugsberechtigten kein finanzieller Nachteil aus der Streichung des Pflegewohngeldes entstünde. Materiell belastet würden lediglich leistungsberechtigte Nicht-Sozialhilfeempfänger dadurch, dass sie

[9] Vgl. § 82 Abs. 4 SGB XII

[10] Vgl. § 1 Abs. 1 Nr. 1 a) BSHG§88Abs2DV 1988 - Verordnung zur Durchführung des § 90 Abs. 2 Nr. 9 SGB XII

mehr, bzw. früher ihr eigenes Barvermögen einbringen müssten, da sie nicht länger über den Anrechnungsfreibetrag von 10.000 Euro Barvermögen verfügen könnten, sondern nur über den geringeren der Sozialhilfe. Typischerweise wird bei ihnen durch das Pflegewohngeld der Bezug von Sozialhilfeleistungen um einige Monate verzögert, nicht aber abgewendet, da die laufenden Kosten der Lebenshaltung ja weiterhin aus eigener Kraft bestritten werden müssen, solange dies möglich ist und Sozialhilfebedarf noch nicht gegeben ist.

Die Gruppe der finanziell Benachteiligten ist a) so klein und b) nur so überschaubar benachteiligt, dass ein gravierender Nachfrageausfall für einen Betreiber wohl nicht zu befürchten ist. Ein bauwilliger Betreiber wird sich dadurch vom Bau einer neuen Einrichtung kaum abbringen lassen. Rechtlich kann ihm weder Bau noch Betrieb untersagt werden. Es steht also zu befürchten, dass die gewünschte Steuerungswirkung, die Bekämpfung von amtlich vermuteten Überkapazitäten durch die Streichung des Pflegewohngeldes, nicht zu erreichen ist.

Überaus zweifelhaft ist darüber hinaus, ob diese landesgesetzliche Regelung mit der Berufsfreiheit in Art. 12 GG vereinbar ist. Juristisch bestehen erhebliche Zweifel an der Rechtmäßigkeit des Landesgesetzes bzw. seiner Anwendung. Bereits 2001 urteilte das Bundessozialgericht über ein ähnlich gelagertes Landesgesetz in Rheinland-Pfalz und stellte klar, dass es den Ländern „im Hinblick auf das Grundrecht der Berufsfreiheit und aus dem Gesichtspunkt der Bundestreue untersagt [ist], Pflegeeinrichtungen, die von den Pflegekassen zugelassen sind, als nicht bedarfsgerecht von der finanziellen Förderung auszuschließen“ [8]. Bei einem Überangebot dürfe die Förderung von stationären Einrichtungen nicht für einzelne Leistungsanbieter, sondern nur generell für alle Heime eingestellt werden [9], führt das Bundessozialgericht in seinem Urteil aus:. „Entschließt er [der Landesgesetzgeber, Anm. d. Verf.] sich aber zu einer Förderung, ist diese aus verfassungsrechtlichen Gründen wettbewerbsneutral auszugestalten.“ [9] Der vom Bundesgesetzgeber gewünschte Leistungswettbewerb unter den Leistungserbringern dürfe keinesfalls beeinträchtigt werden [10].

Die Vermutung, dass das Alten- und Pflegegesetz NRW in Hinblick auf die Konsequenzen einer verbindlichen Pflegeplanung verfassungswidrig ist und daher nicht ewig Bestand haben wird, erscheint vor diesem Hintergrund nicht wirklich gewagt. Zumindest aber sei den Kreisverwaltungen geraten, den überschaubaren Einsparmöglichkeiten alle Nachteile gegenüberzustellen. Dazu zählen neben den Nachteilen durch unterbleibende Investitionen auch die Kosten durch mit Sicherheit zu erwartende Rechtsstreitigkeiten.

5. Literatur

[1] BVerfG, Urteil vom 11.06.1958 (BVerfGE 7,377).

[2] BVerfG, Urteil vom 11.06.1958 (BVerfGE 7,377, Leitsatz 6. c).

[3] Landtag Nordrhein-Westfalen (2003): Entwurf eines Gesetzes zur Änderung des Gesetzes zur Umsetzung des Pflegeversicherungsgesetzes (Landespflegegesetz Nordrhein-Westfalen - PfG NW). Drucksache 13/3498, vom 03.02.2003.

[4] Landtag Nordrhein-Westfalen (2003): Drucksache 13/3498, vom 03.02.2003.

[5] Vgl. BeckOK SozR/Siebel-Huffmann SGB XII § 75 Rn. 7.

[6] Grube/Wahrendorf (2014): SGB XII. 5. Auflage (Rn. 54).

[7] Vgl. BeckOK SozR/Siebel-Huffmann SGB XII § 75, Rn. 7.

[8] BSG: Urteil vom 28.06.2001. B 3 P 9/00 R 2. Leitsatz.

[9] BSG: Urteil vom 28.06.2001. B 3 P 9/00 R Randziffer 22b.

[10] BSG: Urteil vom 28.06.2001. B 3 P 9/00 R Randziffer 25.

Prof. Dr. Hanjo Allinger

Hanjo Allinger studierte Volkswirtschaftslehre an der Universität Passau. Ein Stipendium der Harvard-University ermöglichte ihm 1998/1999 einen Forschungsaufenthalt in Boston. 2003 promovierte Allinger an der Universität Passau in Zusammenarbeit mit dem Institut für Arbeitsmarkt- und Berufsforschung der Bundesagentur für Arbeit in Nürnberg mit einer Arbeit zur Einkommensdiskriminierung. 2008 wurde er an der Cologne Business School in Köln zum Professor berufen, seit 2010 hat er an der TH Deggendorf eine Professur für Volkswirtschaftslehre inne. Schwerpunkte seiner Tätigkeit liegen im Bereich der Bildungs- und Gesundheitsökonomik.
Neben seiner Tätigkeit als Professor an der THD ist Allinger Geschäftsführer des Instituts für empirische Wirtschafts- und Sozialforschung INWISO in München. 2012 wurde er in Anerkennung seiner Forschungsleistungen in die Europäische Akademie der Wissenschaften und Künste gewählt.

Hanjo Allinger studied economics at Passau University. A scholarship of Harvard University during the 1998/1999 academic year allowed him to deepen his research in Boston. In 2003, he received his doctorate from Passau University for his dissertation on labor market discrimination that he elaborated in cooperation with the Institute for Employment Research, the research institute of the Nuremberg-based German Federal Employment Agency. In 2008, he was appointed professor at Cologne Business School. Since 2010, he has been holding a professorship for economics at Deggendorf Institute of Technology. The focus of his research lies in the fields of educational economics and health economics. Besides his professorship in Deggendorf, Hanjo Allinger is managing director of INWISO, an institute for empirical economic and social research, located in Munich. In recognition of his research activities, he was elected member of the European Academy of Sciences and Arts in 2012.

Kontakt / *Contact*

✉ **hanjo.allinger@th-deg.de**

BAVARIAN JOURNAL OF APPLIED SCIENCES®

Call for Papers

Issue No. 3 (2017)

The Bavarian Journal of Applied Sciences (BJAS) is a young open access and peer-reviewed journal that is published by Deggendorf Institute of Technology, Bavaria. The BJAS provides an interdisciplinary platform for researchers in applied sciences for imparting and sharing knowledge in the form of empirical and theoretical research papers. The journal welcomes and acknowledges submissions from researchers in management, economics, tourism, health sciences, IT, technology, natural sciences and engineering. It is available in both print and online format.

For the next issue, which will be published in December 2017, BJAS invites papers that contribute to the following topic:

Opportunities and Risks of Digitalization for the Public, Private and Industrial Sectors

Digitalization provides a broad range of opportunities for the private and industrial sectors, including the production process, logistics and transport, but also for the public sector, including administration and management as well as education. Digitalization potentially allows for automatization and for the annihilation of distance. However, due to the need for specialized technological knowledge and skills, digitalization also generates new risks and challenges regarding security, data confidentiality and employment. For its third volume, the BJAS welcomes papers analyzing the opportunities and the risks as well as those providing possible solutions to meet the challenges of digitalization.

Contributions will be chosen for review by the journal's scientific editorial board. They will be reviewed in a double-blind peer review process. Please send your manuscript to the editorial team at **info@jas.bayern**, considering the instructions for authors and the style guidelines which may be found on the journal's website. The deadline for submission is **June 1st, 2017.** Notification of acceptance for review will be given until July 1st, 2017. Papers can be submitted in either English or German. For further information on manuscript submission and the review process please see **www.jas.bayern**.

Bavarian Journal of Applied Sciences
Engineering, Information Technology, Economics, Health Sciences

No. 2 (December 2016)

Principal Editors:
Prof. Dr. rer. nat. Peter Sperber & Prof. Dr.-Ing. Andreas Grzemba.
Deggendorf Institute of Technology, 2016

BAVARIAN JOURNAL OF APPLIED SCIENCES