1. Details of applicant
Name, title(s): Elzinga, Dr. C.H.
Affiliation (incl. fte): M&T, 1 fte
Telephone: 020-5986889
E-mail: ch.elzinga@fsw.vu.nl
Website: (http://home.fsw.vu.nl/ch.elzinga/)

Duration: x Application for an additional 0.2 fte research time during one year

Discipline: Analysis of Categorical Time Series
Keywords: Sequence Analysis, OM, Combinatorial Algorithms

2. Title of the application
Restructuring CHESA – a new implementation of a VU-FSW sequence analysis software package.

3. Research proposal
3a. Problem statement and theoretical background

In many branches of science and engineering, data come in the form of sequences of encoded observations, the sequential character arising as a result of an ordering in time or space. Typical examples from the social sciences are the categorical time series that result from encoding life course or career events. In the case of categorical time series, an event-property like duration may be used to further characterize or detail the separate events.

One way to analyse sets of such sequences is to construct spatial representations of such datasets and use the distributions of distances or specific configurations of points to formulate and test substantive hypotheses. For example, the average distance between representations of life courses of parents and children may be used to study the intergenerational transfer of life course patterns (e.g. Liefbroer and Elzinga, 2007).

In the early eighties, Andrew Abbott and his collaborators introduced in the social sciences a way to construct such metric representations: Optimal Matching (OM). For almost twenty years, OM was the only technique available for constructing distances. In recent years, its application has been seriously criticized (e.g. Settersten and Mayer, 1997; Wu, 2000) for a number of reasons, most particularly regarding the fact that the metric used finds its roots in 2003, I introduced an alternative to OM (Elzinga, 2003) that is based on a combinatorial attribute of pairs of sequences (the number of matching subsequences) and in 2005 (Elzinga, 2005), I adapted the metric to handle durations. Since then, I have proposed (Elzinga, 2007) a range of alternatives to OM, each one based on a distinct combinatorial attribute of pairs of sequences. Each of the proposed metrics is easy to interpret in terms of the phenomena it is applied to and each method handles durations in a straightforward, natural way. In collaboration with Aat Liefbroer, I demonstrated the applicability of some of these metrics; we used them to test hypotheses on de-standardization of family formation of young adults (Elzinga and Liefbroer, 2007) and to model transfer of patterns of family formation across generations (Liefbroer and Elzinga, 2007). Together with Hilde Bras and Aat Liefbroer, I am working on an analysis of life course trajectories that have recently been reconstructed from archives dating back to the late 19th century. The results will be published in a 2008-edition of Mens & Maatschappij. In addition, I am preparing a paper on determinants of (very) early motherhood with Gaia Salford, a Ph.D-student of Francesco Billari from the Bocconi University in Milan and Raffaella Piccarreta. Furthermore, together with Raffaella Piccarreta, associate professor in statistics at the Bocconi-University who is presently at our department on a three month sabbatical, I am preparing a paper on the methodology of relating categorical time series from different
domains like e.g. family formation and the education/work-domain. This newly developed methodology builds on the above metrics.

Finally, the combinatorial algorithms that I developed to numerically evaluate the metrics proposed, are related to more general problems that have applications in kernel-based machine-learning and high-throughput DNA-sequencing. This more general application of these algorithms has lead to a joint paper with Hui Wang from the School of Computing and Mathematics of the University of Ulster and Professor Sven Rahmann of the Institute for Bioinformatics for High-Throughput Technologies of the Technical University of Dortmund.

In order to use each of these proposals, I developed software that essentially calculates the distance matrices according to the proposed metrics. This software, with its manual, has been made publicly available under the name of CHESA, freely and unconditionally downloadable from my VU-homepage. Presently, this software has at least 30 users in Europe and Northern America; a list of known users is attached to this application as Appendix C.

Over the years, this software (some 10.000 lines of C#-code) has grown into a structure that is difficult to maintain or to extend because its architecture is no longer suitable for its present functionality. Therefore, I am applying for this CCSS fellowship to support the partial redesigning and re-implementation of the CHESA-package.

3b. Aspects of the planned re-programming

- Increase the tolerance to irregularities in the input files. The present version is extremely rigid and unforgiving. For ordinary users, this can be very frustrating since it is often not immediately clear to them what causes the problems. This in turn leads to frequent requests for help from first-time users.
- Inclusion of some graphical tools for exploratory analysis of data. To accomplish this, I will buy a license to use a set of plotting primitives and integrate these with CHESA.
- Implementing an aliasing structure for the combinatorial procedures. This will make the process of extending the program, and subsequent testing, faster and easier. Special versions of the program will be easier to generate.

4. Key publications relevant to the present proposal


5. Time Plan

The first activity will focus on improving CHESA’s abilities to endure and absorb minor errors and add some procedures to guide the user when the errors are too serious or frequent. I expect that this can be implemented and tested in 5 to 10 days. Most of this time will be used to thoroughly and systematically test the changes.

Next, I will replace the complicated and redundant structure that leads to the choice of the combinatorial algorithms by an aliasing mechanism. Approximately 10 days are required to complete this task.

The most difficult and time consuming job will be the construction of graphical enhancements because they require the construction of a much more sophisticated output device. This part of the work will take the remaining amount of time.

6. Relevance

For my co-researchers in methodology and myself, having software that is more modular and easier to adapt to new algorithms will create the possibility to focus more on the
substantive and methodological issues of research and less on the implementation, testing and maintenance of the program. For all users of the package, new releases will be more stable, better documented and easier to use and it will be easier to create and maintain special versions made on request for individual users. For students (the methods I proposed are taught by e.g. Brendan Halpin during the Essex Summer School on Social Science Data Analysis & Collection) and researchers, I will include graphical features that will allow for color-renderings of e.g. job-careers or classifications of educational trajectories.

7. Expected output and contributions

7a. Expected output

The expected output will consist of a code and documentation that in comparison with the present release 2.11 will be better structured, more modular, more stable and easier to use and adapt. Although no scientific publication will be the direct result of the work planned, this work will certainly facilitate my writing of scientific papers in the future because it will save a lot of time: the new version will require less maintenance. Users will require less help and it will be much easier to adapt the program to new applications and metrics. Furthermore, since the software will be easier and more attractive to use, it will lead to an increase of the number of citations of my work.

7b. Contribution to the research program

I am a participant in the Comparative Stratification Research program that is directed by Harry Ganzeboom. With Aat Liefbroer and Hilde Bras, both members of the same research program I am currently working on a project involving the analysis of the recently published dataset “Historische Steekproef Nederland”, a dataset that contains life course data of Dutch people born between 1850 and 1920. A part of this analysis will be published in a special issue of Mens & Maatschappij. Test-versions of the new CHESA-program will be used for this purpose. Another project with Irma Reci, also a participant of CSR, on job/salary careers is planned for upcoming January. More applications in the context of CSR-research are to be expected in the years to come.

7c. Contribution to the research mission of the CCSS

A central theme of the CSR is the study of educational and professional mobility in relation to various aspects of the life course (e.g. education, work and family formation, income). Therefore, developing adequate methodology to study and model categorical time series is vital to the program. Without the development of new methods and algorithms for sequence analysis, it is difficult to investigate scientific questions that are not confined to single life course events but, instead, pertain to life course trajectories as the unit of research.

7d. Additional value

The design and implementation of robust software with a gentle user interface is extremely time consuming and will most probably never be fully carried out with rigor and the required testing if this work will have to be done during regular work time. Therefore, it is necessary to reserve separate time to be able to develop and test this software. This will leave me sufficient time during regular work hours to continue working on my publications and fulfill the faculty’s publishing requirements. No other support for the proposed project has been obtained nor has such support been asked for.

8. Replacement of applicant’s teaching obligations

Replacement of the applicant’s teaching obligations is stated in the attached letter by the Head of the Dept. of Social Research Methods in Appendix A.

9. CV of applicant

My CV has been added as Appendix B to this application.
10. I hereby declare that I have completed this form truthfully:

Name: Cees H. Elzinga

Amsterdam, Date: November 12, 2007

11. Literature references

Appendix B: Cees H. Elzinga’s concise CV

Born in Hilversum (The Netherlands) on January 18, 1950

Assistant professor at the Sub-Faculty of Psychology of the Catholic University of Nijmegen, Department of Mathematical Psychology, 1978-1985.

Publications until 1987


Ph.D in 1985 at the Catholic University of Nijmegen. Thesis: ”On the Measurement of Color and Brightness”; Supervisors: Prof. Dr. E.E.Ch.I. Roskam en Prof. Dr. Ch.M.M. de Weert.


Assistant Professor at the Faculty of Social Sciences of the VU University, 2002 –

Publications since 2002


Software

- CHESA Sequence Analyzer, downloadable from [http://home.fsw.vu.nl/ch.elzinga/](http://home.fsw.vu.nl/ch.elzinga/).

Lectures (invited, expenses paid)

- ”CSA, Practical Software for the Analysis of Job-Careers” to the Interfacultair Instituut WAV in August 2005 in Leuven.
- ”Metric Representation of Categorical Time Series” to the Vereniging voor Ordinatie en Classificatie (VOC) on November 3, 2006 in Leiden.
- ”Transfer of Family Life Trajectories” to ESRC Centre for Research in Socio-Cultural Change (CRESC) on September 3, 2007 in Manchester.
- ”Metric Sequence Representations” to the Wissenschaftszentrum Berlin (WZB) on September 10, 2007 in Berlin.
Submitted


Guests

- June-August 2007: Gaia Salford, Ph.D-student at the Instituto di Metodi Quantitativi of the Bocconi-university in Milan.
- September-December 2007: Dr. Raffaella Piccarreta, Associate Professor of the Instituto di Metodi Quantitativi of the Bocconi-university in Milan.
Appendix C: Known CHESA-users.

I regularly inform known users about updates of the software and/or the manual through e-mail and then use the list below.

jbruederl@sowi.uni-mannheim.de
leen.heylen@ua.ac.be
dimitri.mortelmans@ua.ac.be
trevor.murrells@kcl.ac.uk
huangqh@psychology.su.se
cwilson@cmhc-schl.gc.ca
andrew.harvey@stmarys.ca
leov@staff.ttu.ee
brendan.halpin@ul.ie
gindo.tampubolon@manchester.ac.uk
stovel@u.washington.edu
jacques-antoine.gauthier@unil.ch
d.mcvicar@qub.ac.uk
roorjm@fsw.leidenuniv.nl
a-abbott@uchicago.edu
Tom.Schatteman@ua.ac.be
stefan.iske@uni-bielefeld.de
ornalior5@hotmail.com
raffaella.piccarreta@uni-bocconi.it
g.Pollock@mmu.ac.uk
gaia.salford@phd_unibocconi.it
brzinsky-fay@wz-berlin.de
aciqg@kcl.ac.uk
A.Manzonii@uvt.nl
a.fasang@iu-bremen.de
chung.ip@nuffield.oxford.ac.uk
luniak@wzb.eu
kohler@wzb.eu
laurent.lesnard@sciences-po.fr
brzinsky-fay@wzb.eu

Since CHESA is freely downloadable from my homepage http://home.fsw.vu.nl/ch.elzinga/, there might be more users.