Agenda EMOS symposium 02:00 PM – 05:00 PM CEST via Zoom

1. Welcome and opening symposium – by chair Reinoud Stoel
2. Introduction to the EMOS Network and EMOS activities – by Carola Carstens and Tina Steenvoorden (EMOS Secretariat at Eurostat)
3. EMOS at CBS – by Jacco Daalmans (CBS)
4. Presentation 1 – by Eulàlia Gómez-Aguiló (Leiden University):

From noisy data to sound conclusions: Testing the limits of contagion models

Description: Behaviors and ideas, from smoking and diet choices to technological innovation, spread through societies much like infectious diseases, making contagion models a powerful lens for understanding social and economic dynamics. Yet, unlike carefully monitored health data, the observational data behind these processes are often messy, incomplete, and delayed, raising questions about how much we can trust the conclusions drawn from them. This thesis tests the robustness of epidemic models by systematically distorting real COVID-19 data from the Netherlands to reveal how data imperfections shape the reliability of inferred results.

5. Presentation 2 – by Leonie Abrahams (Leiden University):

Bias and representativity in linked administrative data

Description: Linking data from different sources is more complex than it may appear. When no personal identifier is available, you must choose which characteristics to use to match individuals across datasets. But what happens when several people share the same characteristics, or when the information differs between sources? This project examines different methods for linking data and investigates how these choices affect the final data set. It also explores ways to measure representativeness of population subgroups after linkage, and how linkage decisions may introduce bias.

6. Presentation 3 – by Robert van der Kaap (Utrecht University):

Statistical Matching with a Proxy Variable: The Role of Measurement Error, Sample Size, Overlap, Selectivity, Proxy Quality and the Validity of the Conditional Independence Assumption.

Description: We often combine datasets to answer important questions, but how trustworthy are the results? This presentation explores when these methods work, when they don't, and why data quality matters.

7. Presentation 4 – by Mila Madiot (Leiden University):

The model selection process for estimating the number of homeless people in the Netherlands

Description: The estimation of the number of homeless people in the Netherlands uses a methodology inspired by ecological statistics, called Capture-Recapture. Many ways have been developed to obtain estimates of population sizes, but the stability of the results across methods has only partially been assessed. Hence the question: how stable are the estimates of the size of the homeless population across different model selection processes?

Break (15 min) 15:30-15:45

8. Presentation 5 – by Reza van Welie (Leiden University):

A breadcrumb trail through the forest
Understanding machine learning imputations for educational attainment

Description: Machine learning models are beneficial for official statistics, but their black-box nature limits trust and adoption. This presentation demonstrates how post hoc interpretability methods can unlock these models and bring transparency to imputing missing educational attainment.

9. Presentation 6 – by Milena Costa (Leiden University):

The Contexts that Bind Us

Description: People are connected in many different ways, but how much do these connections overlap? Using Dutch population registers covering over 17 million residents per year, I develop a measure that captures when relationships span multiple institutional domains: when a classmate is also a colleague, or family, school, and neighbourhood ties intersect. I study how this overlap varies across regions and social groups, and what it might mean for social cohesion and trust.

10. Presentation 7 – by Josep Ferrer (Leiden University):

From Voting Booth to Nationwide: Measuring and Spatially Interpolating Electoral Polarization in the Netherlands

Description: Political polarization. A term that has been talked about in the news incessantly over the last decade. It would seem as if this process is everywhere, and it would be logical that how prevalent this phenomenon is would be more researched. However, this is not the case. Dutch politics are no strangers to talk of polarization, as the elections in 2025 brought about a marked split between right and left parties that seemingly now do not want to work with one another. From here one could go and ask if this divide is also present in the

population, not necessarily based on beliefs, but on where they are located. In other words, can this political polarization be located, measured and mapped?

11. Presentation 8 – by Shangheng Teng (Leiden University):

Classifying Text into Hierarchical Taxonomies: An Active Learning-Assisted Meta-Analysis of NLP Performance

Description: Official statistical offices are increasingly using AI and Natural Language Processing (NLP) to automatically classify texts, such as job descriptions or product names. But how accurate and reliable are these models, really? In my presentation, I will show how we evaluate the "true" performance of various NLP models to help CBS choose the best tools for complex classification tasks.

12. Presentation 9 – by Marit Rouwenhorst (Leiden University):

Exploring improving physical activity estimates based on a survey with accelerometer data.

Description: In this presentation, I will discuss the way Statistics Netherlands is measuring people's physical activity currently, using a survey, and if we can improve this using data from an accelerometer, a physical activity sensor. Explaining how the outcomes of the survey and the accelerometer differ and if we can use the benefits of both to try to get an even better idea of how active people are.