Nederlandstalige samenvatting CPB Discussion Paper 441

'Forecasting World Trade Using Big Data and Machine Learning Techniques'

'Het voorspellen van de wereldhandel met big data en machine learning'

Voor kleine, open economieën zoals Nederland is het belangrijk om te weten hoe de wereldhandel zich ontwikkelt, omdat fluctuaties in de wereldhandel sterk kunnen doorwerken op de ontwikkeling van de economie. Voor beleidsmakers in Nederland is het daarom van belang om te weten hoe de wereldhandel zich in de toekomst gaat ontwikkelen. Voorspellingen zijn hiervoor noodzakelijk.

Het huidige voorspelmodel voor de wereldhandel van het CPB, de BVAR, maakt gebruik van een set van 23 economische variabelen om de wereldhandel te voorspellen. Gebruikte voorspellers zijn bijvoorbeeld de omvang van de wereldhandel in het verleden, de olieprijs en productie van staal of halfgeleiders. Deze variabelen worden gecombineerd in een groot Bayesiaanse Vector AutoRegression-model (BVAR) dat ontwikkelingen in deze variabelen over de tijd gebruikt om de toekomstige wereldhandel te voorspellen. Deze *core*-dataset begint in maart 1991 en eindigt in april 2020.

De omvang van verklarende Bayesiaanse VAR-modellen is beperkt tot enkele honderden variabelen. Hoewel dit aantal relatief groot is, vereist de aanwezigheid van vele duizenden economische variabelen dat economisch experts op voorhand kiezen welke variabelen opgenomen worden in dit model. Het risico bestaat dat experts niet alle relevante voorspellers voor de wereldhandel opgenomen hebben in het model. Daarom onderzoeken we in dit paper of voorspellingen van de wereldhandel verbeteren als we gebruik maken van *big data* en *machine learning*. Onze *big* dataset bestaat uit alle maandelijkse economische indicatoren die beschikbaar zijn op Datastream tussen maart 1991 en oktober 2019, aangevuld met enkele tijdreeksen die volgens economisch experts van het CPB nuttig zijn. In totaal zitten er 11.017 tijdreeksen in de *big* dataset.

Machine learning-modellen zijn nodig om wereldhandel te voorspellen op basis van deze *big* dataset. Het aantal verklarende variabelen is immers fors groter dan de ongeveer 400 maanden waarvoor de omvang van de wereldhandel gemeten is. Hierdoor kunnen traditionele econometrische modellen, zoals de BVAR, geen voorspellingen doen op basis van de *big* dataset. *Machine learning*-modellen kunnen dat wel, omdat zij zelf op basis van de data bepalen welke variabelen gebruikt worden. Hierdoor kan het aantal gebruikte variabelen veel groter zijn dan het aantal waarnemingen.

In het paper gebruiken we de meest bekende *machine learning*-technieken. We voorspellen wereldhandel met een LASSO, met een Random Forest en met een ensemble van deze modellen en de BVAR geschat op de *core data*. Deze methoden gebruiken de data op verschillende manieren. LASSO selecteert bepaalde variabelen en zoekt een lineair verband tussen deze variabelen en de toekomstige wereldhandel. Een Random Forest maakt juist een niet lineaire combinatie van voorspellers, maar kan dit in iedere stap slechts doen voor een beperkt aantal (willekeurig bepaalde) voorspellers. LASSO en Random Forest gebruiken de meest recent gepubliceerde observatie van alle variabelen en drie vertragingen van het

eerste verschil. Het ensemble combineert de voorspellingen van LASSO, Random Forest en de BVAR door het gemiddelde van deze voorspellers te bepalen.

We gebruiken deze *machine learning*-modellen en de BVAR om de toekomstige wereldhandel te voorspellen. De machine learning-modellen doen dit op basis van de *core*-dataset en de *big* dataset, terwijl de BVAR noodzakelijk enkel de *core* dataset gebruikt. We voorspellen de wereldhandel 1, 2, 3 en 6 maanden vooruit. Omdat de publicatie van de wereldhandel een vertraging heeft van ongeveer twee maanden, is de voorspelling twee maanden vooruit een zogeheten *nowcast*.

We vergelijken de voorspellingen van de wereldhandel van deze modellen voor de periode die begint in april 2006. We gebruiken hierbij een *rolling out of sample RMSE*. Dus voor de voorspelling 1 maand vooruit voor april 2006 gebruiken de modellen alle gegevens uit de datasets die gepubliceerd zijn tot en met maart 2006. Voor de voorspelling voor mei 2006 gebruiken de modellen alle gegevens in de datasets die gepubliceerd zijn tot en met april 2006, etc. Vervolgens wordt de RMSE (root mean squared error) van al deze (*out of sample*) voorspellingen bepaald. Om te kijken hoe de RMSE verandert over tijd, bepalen we deze voor een window van 13 opeenvolgende voorspellingen. Dit maakt het mogelijk om na te gaan hoe de voorspelkracht van modellen wijzigt over tijd. We gaan hier specifiek in op de voorspellingen tijdens de grote financiële crisis (augustus 2008 tot en met augustus 2009) en de perioden daarbuiten.

Onze resultaten laten zien dat de *machine learning*-modellen op basis van *core data* vergelijkbaar presteren als de BVAR (geschat op dezelfde data). Tijdens de financiële crisis voorspellen de *machine learning*-modellen slechter dan de BVAR, erbuiten voorspellen ze beter. Dit patroon wordt iets sterker als de *machine learning*-modellen toegang hebben tot de *big data* en gaat op voor alle onderzochte voorspelhorizons. De RMSE van de *machine learning*-modellen wijkt echter niet significant af van de RMSE van de BVAR, wanneer we corrigeren voor het feit dat we een groot aantal hypothesen testen volgens de Holm-Bonferroni-correctie.

De voorspellingen van de wereldhandel met *machine learning* op basis van de *core* of *big data* zijn vergelijkbaar met die van de BVAR (geschat op de *core data*). Aanvullende analyses doen vermoeden dat dit komt doordat de *machine learning*-modellen '*overfitten*': er zijn te weinig observaties om nauwkeurig te bepalen hoe relevant specifieke voorspellers in de data zijn. Hoewel sommige papers concluderen dat *machine learning*-modellen bepaalde macro-economische tijdreeksen beter voorspellen dan conventionele econometrische modellen, geven onze resultaten aan dat dit niet zondermeer opgaat wanneer de wereldhandel voorspeld wordt met een grote Bayesiaanse VAR.