Exploring Alternative Data for Nowcasting: A case study on US GDP using Topic Attention

Lucas Manchado Marcos, Ariel Duarte-López, **Argimiro Arratia**
Universitat Politécnica de Catalunya & Acuity Trading S.L. Barcelona
ECML-PKDD+ MIDAS 2023

Nowcasting & Alternative Data

NOWCASTING is the estimation of **current** or near-term economic conditions in real-time or with **minimal time lag.**

ALTERNATIVE DATA refers to **non-traditional** or unconventional data sources that are used to supplement or complement traditional data sets in the analysis and decision-making process.

Input Data - Gross Domestic Product

The official source is the **Bureau of Economic Analysis** (BEA).

Different estimations based on surveys or economy indicators release (quarterly): **Advanced**, **Second** and **Third** estimation.

Target variable is the **US GDP growth**.

Input Data - Economic Indicators

The chosen indicators are motivated by the works of *Hopp (2022b)*.

The official source is the **Federal Reserve of Economic Data** (FRED)

Aggregation bias problem or **mixed frequency** models.

Variable Label	Description	Release
ULCNFB	Nonfarm Business Sector: Unit Labor Costs for All Workers.	Quarterly
A261RX1Q020SBEA	Real Gross Domestic Income.	Quarterly
PAYEMS	Number of US workers.	Monthly
CPIAUCSL	Consumer Price Index for All Urban Consumers.	Monthly
UNRATE	Unemployment Rate.	Monthly
HOUST	New Privately-Owned Housing Units Started.	Monthly
INDPRO	Industrial Production.	Monthly
CPILFESL	Consumer Price Index for All Urban Consumers.	Monthly
DSPIC96	Real Disposable Personal Income.	Monthly
PCEPILFE	Personal Consumption Expenditures Excluding Food and Energy.	Monthly
PCEPI	Personal Consumption Expenditures.	Monthly
PERMIT	New Privately-Owned Housing Units Authorized in Permit-Issuing Places.	Monthly

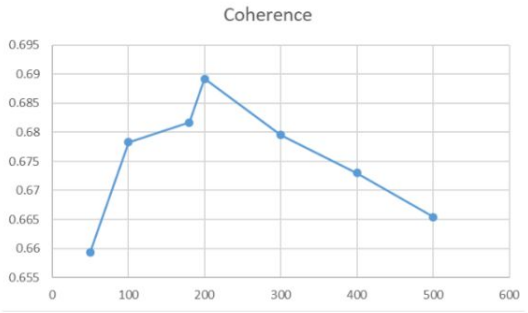
Hopp, D. (2022b). Economic nowcasting with long short-term memory artificial neural networks (LSTM). Journal of Official Statistics, 38(3):847–873.

Input Data - Topic Model

Latent Dirichlet Allocation (LDA) model trained by Acuity Trading S.L..

The training dataset is a corpus of over 2 MM financial news spanning from 2001 to 2021.

Choose the optimal number of topics using a **coherence** measure (calculates the semantic similarity between high-scoring words within the topics.)

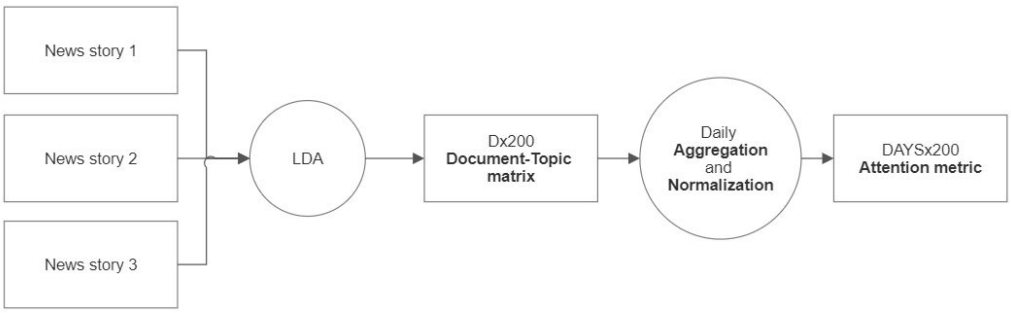


Input Data - Attention Metric

Only keep news stories with more than **100 words** and with more than **10%** of the content captured by one topic.

Daily attention metric is the sum of the probabilities associated with each topic throughout the day and **normalizing** considering the **count** of all processed news stories in that day.

Metric is **interpretable** since its value ranges between 0 and 1.

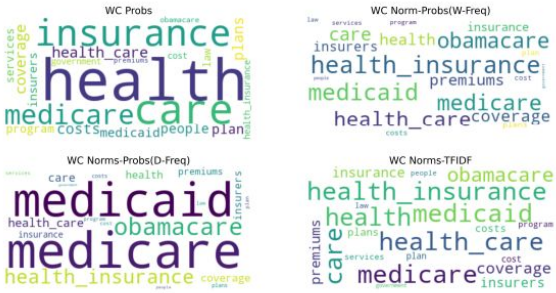


Input Data - Topic Labelling

The **Topic-Term** matrix contains, for every topic, the probability distribution of every word in the corpus.

200xN
Topic-Term
matrix

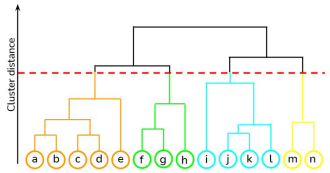
Create **word clouds** using different normalization strategies.



Labelling is manual (e.g. Healthcare)

Descriptive Analysis - Cluster Generation

Use of an **Agglomerative Hierarchical Clustering** algorithm, which organises similar topics based on the weight assigned to each word, from the **Topic-Term** matrix.



Two important parameters:

- **Affinity** defines the metric used to compute the linkage.
- **Linkage** determines the distance metric used between sets of observations.

End up with **66** attention clusters

Generate a **word cloud** of each cluster using the mean of the weights of every word from every topic in the cluster to label it.

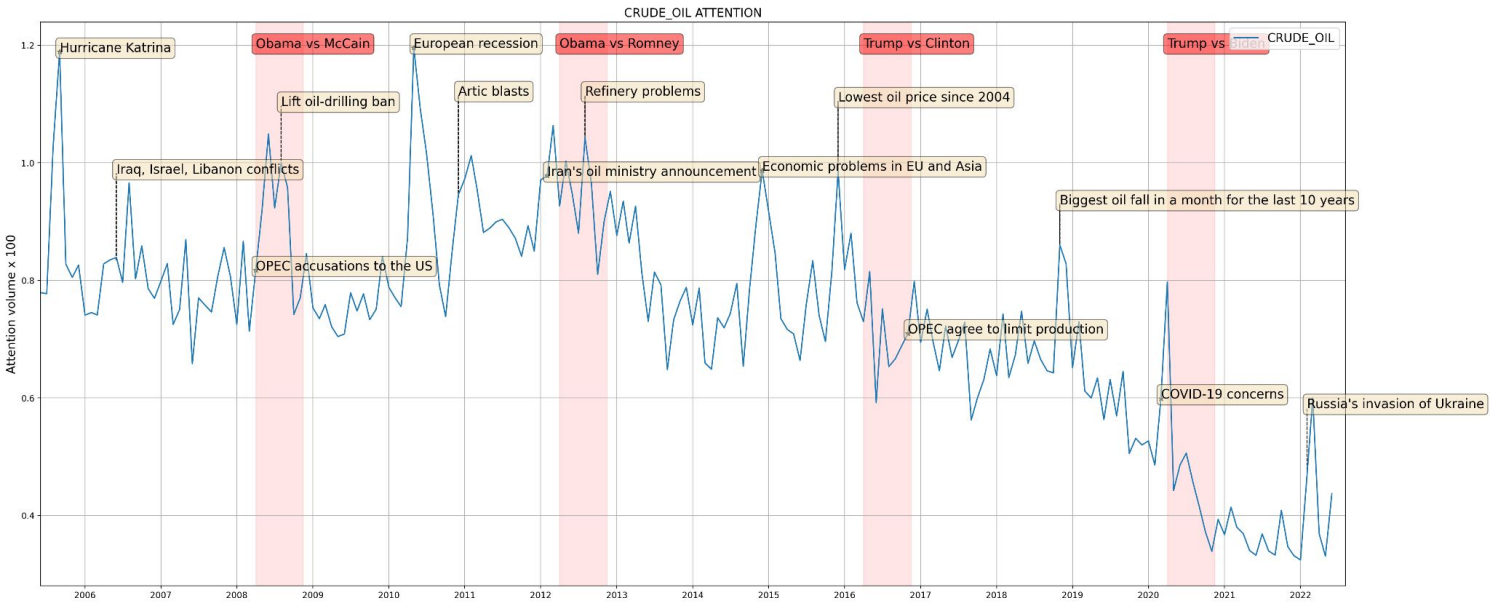
Cluster: Crude Oil

Descriptive Analysis - Cluster Series Analysis

Visual analysis used to **verify** whether the attention clusters **accurately capture** the level of media recognition for historical events.

Monthly aggregated attention clusters for visual clarity.

Election periods are highlighted.



Descriptive Analysis - Quantitative Analysis

Analyze the **relationship** between clusters of the attention metric and the US GDP growth.

Correlation: Distance Correlation, Spearman's Rank Correlation

Causality: Granger Causality

The attention value is aggregated to **quarterly sampling** for consistency.

Moving **windows** approach.

Descriptive Analysis - Correlation

Distance correlation and **Spearman's** Rank correlation.

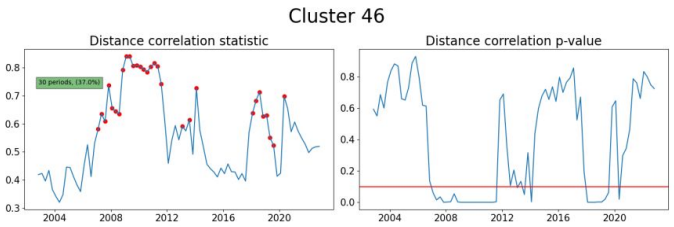
Detect **linear and nonlinear** dependencies.

Null hypothesis assumes **independence**.

Significance level for the test is **0.1**.

Count the number of significant windows to determine the clusters with the **highest correlation.**

Clusters	Nbr. of correlated periods	Topics in Cluster
Cluster 46	30	7 Payments
Cluster 17	28	103 US presidents 189 Life Science Research
Cluster 31	27	45 Company Board 98 SEC 113 Compensation 158 Accounting Scandals



Descriptive Analysis - Granger Causality

Evaluates whether time series X (attention of a cluster) can be used to **forecast** ts Y (US GDP).

Must be applied **bidirectionally**

H0 : X is **not** the cause of Y with a certain delay.

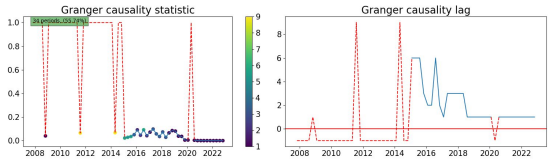
H1: X is the cause of Y with a certain delay..

Significance level for the test is **0.1**.

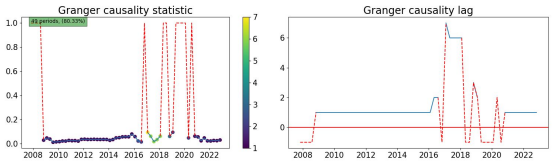
Count the number of significant windows to determine the clusters with the **highest causality.**

Clusters	Nbr. of periods where the cluster attention Granger causes US GDP percent change	Nbr. of periods where US GDP percent change Granger causes the cluster attention	Topics in Cluster
Cluster 9	58	22	38 Gender relations 123 Healthcare 142 Diseases 186 Cancer Research
Cluster 24	49	34	79 Treasury Secretary
Cluster 39	48	14	5 Labor 144 Employment 2

GDP Granger-Causes Cluster 24



Cluster 24 Granger-Causes GDP



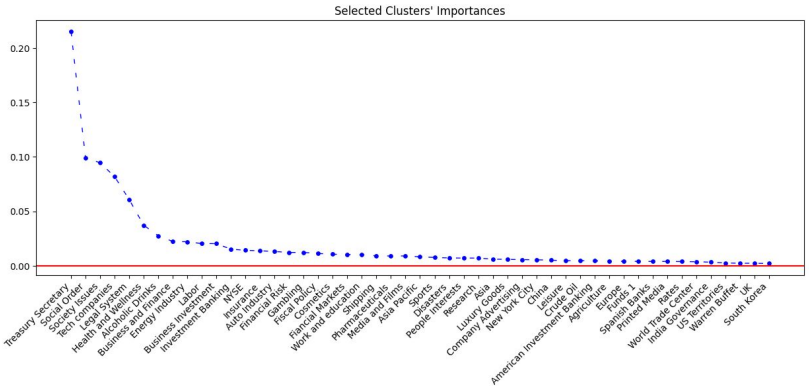
Descriptive Analysis - Discourse Evolution

Identify the clusters that have the **greatest impact** on the US GDP growth.

Use of **Recursive Feature Elimination** (RFE) applied to a **Random Forest** (RF) Regressor.

Train from 2000 to 2018, **validation** from 2019 to 2021 and **test** on 2022.

Validation is necessary for hyperparameter **fine-tuning**.



Methodologies

Long Short Term Memory (LSTM) network

- Special type of **Recurrent Neural Network** (RNN) that can allow outputs from some nodes to affect future inputs for the same nodes.

Mixed Data Sampling (MIDAS) model

- Solves the challenge of incorporating time series data sampled at **different frequencies** into a single model by incorporating the high frequency variables in the form of lags.
- It is a kind of **distributed lag model**.

Methodologies - LSTM

Limitations of RNNs are the **exploding gradient** and the **vanishing gradient***.

LSTMs use the **memory cell** instead of the ordinary recurrent nodes present in RNNs.

The memory cell computes two **states**:

- **Internal** state

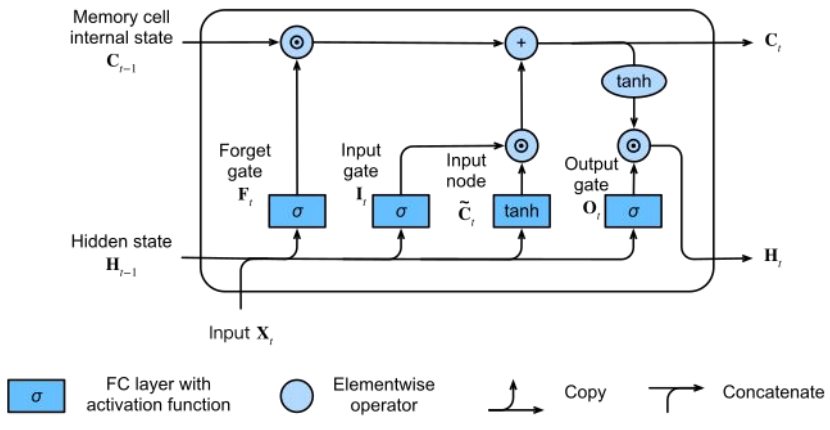
$$C_t = F_t \odot C_{t-1} + I_t \odot \tilde{C}_t$$

- **Hidden** state

$$H_t = O_t \odot \tanh(C_t)$$

The memory cell contains several **gates**:

- **Input** gate
- **Forget** gate
- **Output** gate



Methodologies - MIDAS

Impose a **parametric restriction** on the lag polynomial and add an **autoregressive term** following *Clements and Galvão (2008).*

Restricted MIDAS equation is as follows:

$$y_t = \beta_0 + \sum_{i=1}^{i=3} \alpha_i y_{t-i} + B(L^{1/m}; \theta) x_t^{(m)} + \epsilon_t^{(m)}$$

The estimated parameters are now **3 + R**, where R are the lag distribution parameters for every variable.

10 variables with 10 lags
- Unrestricted estimation → 100 parameters
- Restricted estimation: 3 hyperparameters → 30 parameters

This model also allows to add as many lags in the model as we want, since the number of estimated parameters are **independent** of it.

Clements, M. P. and Galvão, A. B. (2008). Macroeconomic forecasting with mixed-frequency data: Forecasting output growth in the United States. Journal of Business & Economic Statistics, 26(4):546–554.

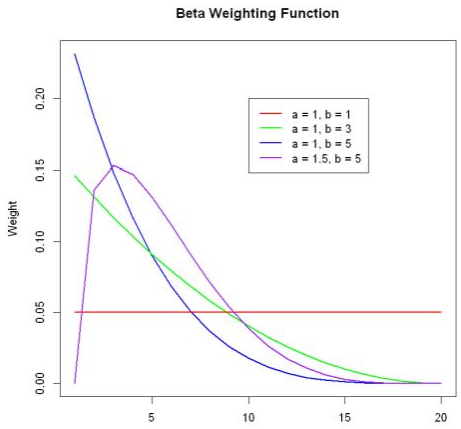
Methodologies - MIDAS

The function used to restrict the lag polynomial is the **Beta probability density function**.

The normalized Beta function is parameterized by $\theta = (\delta, a, b)$

$$B(j;\theta) = \delta * \frac{f(\frac{j}{j^{(max)}}, a, b)}{\sum_{j=0}^{j^{(max)}} f(\frac{j}{j^{(max)}}, a, b)},$$

δ is the **normalization** parameter while *a* and *b* modify the **behaviour** of the function.



Results

Compare the two models in **different scenarios.**

The RFE explained in the Discourse Evolution section is used to do **variable selection** of the clusters.

Following *Hopp (2022a)*, a single restricted MIDAS model is trained for every variable and then by using a weighting scheme an **average forecast** is performed.

Train, validation and test split is required for both models.

For the LSTM a grid search is performed to find the optimal **hyperparameters** for the network.

The MIDAS model also requires a grid search for the optimal **lag length** and **weighting scheme** of the models of every individual variable.

Hopp, D. (2022a). Benchmarking Econometric and Machine Learning Methodologies in Nowcasting. arXiv preprint arXiv:2205.03318.

Results - Full dataset approach

Monthly aggregated data.

A model **without the attention** is trained for both approaches.

- **Train**: 2000-2018
- **Validation**: 2019-2021
- **Test**: 2022

Performance is assessed based on the estimations of the last quarter of the test set (**2022 Q4**) and **Root Mean Square Error** (RMSE) on the test set.

Model	Variables	Validation RMSE	Test RMSE	Advanced	Second	Third
LSTM_m	Econ. & Att.	0.0095	0.021	$(-0.51, 4.88)$ **2.17**	$(2.14, 3.29)$ **2.33**	$(2.14, 3.2)$ **2.55**
	Econ.	0.0222	0.0247	$(-3.01, 7.14)$ 2.04	$(0.52, 3.58)$ 2.03	$(1.88, 2.28)$ 2.07
MIDAS_m	Econ. & Att.	0.0204	0.0302	$(2.41, 3.43)$ 2.50	$(2.42, 3.39)$ 2.46	$(2.27, 3.14)$ 2.41
	Econ.	0.0211	0.0234	$(1.04, 1.89)$ 1.41	$(1.09, 1.93)$ 1.41	$(0.06, 1.52)$ 1.48
MIDAS_d	Econ. & Att.	0.0517	0.0223	$(2.28, 3.14)$ 2.30	$(2.29, 3.2)$ 2.46	$(2.24, 3.15)$ 2.29
BEA Forecasts				2.9	2.7	2.6

	date	payems	cpiaucsl	unrate	houst	indpro	dspic96	cpilfesl	pcepilfe	pcepi	permit	ulcnfb	a261rx1q020sbea	cluster_3	cluster_4	gdp_pct
0	2000-01-01	0.001751	0.002945	0.000000	-0.042130	-0.000730	0.006541	0.003339	0.002364	0.002693	0.026128	NaN	NaN	NaN	NaN	NaN
1	2000-02-01	0.000847	0.004110	0.020000	0.061698	0.003361	0.003878	0.000555	0.001469	0.003095	-0.020255	NaN	NaN	-0.000503	-0.000777	NaN
2	2000-03-01	0.003699	0.005848	-0.019608	-0.076525	0.003856	0.002686	0.003326	0.002022	0.004246	-0.024217	0.038419	0.019790	0.002117	0.000417	0.015
3	2000-04-01	0.002120	-0.000581	-0.040000	0.013707	0.006343	0.005122	0.001657	0.000652	-0.000825	-0.032688	NaN	NaN	-0.000193	-0.000705	NaN
4	2000-05-01	0.001683	0.001745	0.041667	-0.031346	0.002709	0.004140	0.002206	0.001058	0.000826	-0.033792	NaN	NaN	-0.000508	0.000837	NaN
5	2000-06-01	-0.000348	0.005807	0.000000	-0.010152	0.000745	0.002749	0.002201	0.000749	0.003352	0.018782	-0.016142	0.006631	-0.000096	0.001027	0.075

Results - Full dataset approach

Data from the attention is kept in **daily** sampling.

The MIDAS requires the dataset to be **complete.**

Convert the daily frequency into **trading days**(65 days per quarter).

Remove the **weekend** days by aggregating the data to next monday and removing **20 random observations**.

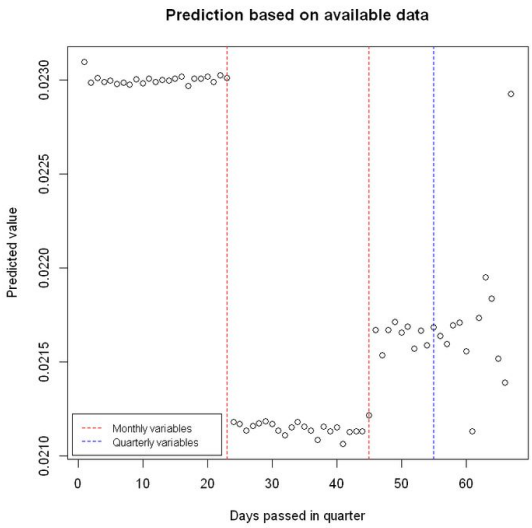
Model	Variables	Validation RMSE	Test RMSE	Advanced	Second	Third
LSTM_m	Econ. & Att.	0.0095	0.021	$(-0.51, 4.88)$ **2.17**	$(2.14, 3.29)$ **2.33**	$(2.14, 3.2)$ **2.55**
	Econ.	0.0222	0.0247	$(-3.01, 7.14)$ 2.04	$(0.52, 3.58)$ 2.03	$(1.88, 2.28)$ 2.07
MIDAS_m	Econ. & Att.	0.0204	0.0302	$(2.41, 3.43)$ 2.50	$(2.42, 3.39)$ 2.46	$(2.27, 3.14)$ 2.41
	Econ.	0.0211	0.0234	$(1.04, 1.89)$ 1.41	$(1.09, 1.93)$ 1.41	$(0.06, 1.52)$ 1.48
MIDAS_d	Econ. & Att.	0.0517	0.0223	$(2.28, 3.14)$ 2.30	$(2.29, 3.2)$ 2.46	$(2.24, 3.15)$ 2.29
BEA Forecasts				2.9	2.7	2.6

date	dspic96	cpilfesl	pcepilfe	pcepi	permit	ulcnfb	a261rx1q020sbea	cluster_5	cluster_6	gdp_pct
<date>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
2000-03-30	NA	NA	NA	NA	NA	NA	NA	0.0004869486	0.0038888887	NA
2000-03-31	0.002685583	0.003325942	0.002022045	0.004246257	-0.02421737	0.03841913	0.01978981	-0.0026146864	-0.0034554048	0.015
2000-04-01	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2000-04-02	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2000-04-03	NA	NA	NA	NA	NA	NA	NA	0.0029266620	0.0019086970	NA
2000-04-04	NA	NA	NA	NA	NA	NA	NA	-0.0007993348	-0.0024963870	NA
2000-04-05	NA	NA	NA	NA	NA	NA	NA	-0.0007301776	-0.0006774043	NA
2000-04-06	NA	NA	NA	NA	NA	NA	NA	0.0019500439	0.0010084198	NA

Results - Full dataset approach

For the daily MIDAS model we can see how the **daily prediction changes** by "hiding" the data that should not be available in that day for the model.

Monthly and quarterly data are incorporated in different moments to understand their **impact**.



Conclusions and future work

Proof of concept that an attention metric is useful to nowcast the value of an economical indicator like the US GDP growth.

The LSTM model generally **performs better** than the MIDAS model, this is aligned with the findings of Hopp (2022a).

Use the attention metric to **nowcast other economic variables**. E.g. Consumer Confidence Index, Unemployment Rate, etc.

Train a **daily LSTM** model to see whether it still outperforms the MIDAS model in that scenario.

Change the average forecast approach used in the MIDAS models and include **all the variables** in the same model

Thank you!