# Predicting irregular migration by online search behaviour

# Research proposal

Joop Adema*

March 15, 2021

**Abstract**

In this research proposal, I suggest the use of Recurrent Neural Networks to predict Illegal Border Crossings (IBCs) at the European Union's outer borders and show the feasibility of such a procedure. As the recent migrant crisis posed a political and humanitarian puzzle for European policy makers and citizens, the ability to forecast irregular migration patterns is important to guide policy makers and to alleviate. Origin country push-factors and host country pull-factors may both be reflected in online search behaviour, potentially enabling improved prediction of future illegal border crossings over sea and land.

Besides the focus on irregular migration and accompanying IBCs, we expand previous work in various directions:

- We use a versatile script to query Google Trend Indices alongside Google Translate for a set of countries and languages that allows for upscaling to source Trend Indices for many keywords.
- We use Google Translate to also include search queries in the local language and don't exclusively rely on a set of English, Spanish or French keywords as in previous work.
- We focus on temporally richer migration data and compare excess performance to a model including fixed effects at various levels.

*adema@ifo.de

As a proof of principle we show that inclusion of several strictly migration-related key words enhances a naive prediction model considerably. Future extensions towards a full fledged predictive model will include closely connected technical, methodological and policy relevant extensions. Among others, those comprise the use of interpretable machine learning to identify important (combinations of) keywords and identify the time scales most strongly relating search behaviour to IBCs.

**Keywords:** Migration intentions; Migration; Internet access; Search behaviour; LIME; G-TAB; RNNs; LSTM

**JEL Codes:** F22

# 1 Introduction

The European Union is a popular destination for migrants of all sorts, including refugees and economic migrants who may have no other means then to migrate illegally. In the prime year of the migrant crises (2015), around 1.8 million illegal border crossings were registred by Frontex.

*Frontières extérieures* (Frontex) is an organization mandated by the European Union founded to primarily coordinate border control on the EU's external borders. Frontex currently organizes several missions on the outer borders where many Illegal Border Crossings (IBCs) take place. After the Syrian civil war broke out in 2011, the number of IBCs sharply rose, peaking in 2015. Although migrant flows and thus IBCs decreased recently, new humanitarian crises could lead to a renewed influx of refugees increasing the pressure on the EU's outer borders and its asylum capacities[1].

Data on Illegal Border Crossings per country of origin and route are provided by Frontex on a monthly basis since 2009, reported with approximately a two month delay[2][3]. Illegal border crossing is one of the ways to obtain the status of an irregular migrant, which also includes entry with illegal documents, overstaying a visa or loss of legal status[4].

On the peak of the migrant crises (in 2015), a yearly number of 1.3 million asylum applications were made in Europe, and 1.8 million IBCs were detected[5]. To put this into perspective, there were only 230 thousand asylum applications in Europe and 114 thousand IBCs in 2009. Over the past 12 years, asylum applications and IBCs were strongly correlated, as many asylum seekers illegally cross the borders before filing for asylum.

[1]For an overview of the current situation, see: `https://frontex.europa.eu/media-centre/news/news-release/situation-at-eu-external-borders-arrivals-down-in-the-mediterranean-2FTL79`

[2]The visualized data and the definition of the various routes can be found here: `https://frontex.europa.eu/we-know/migratory-map/`.

[3]Please note that the number of detected illegal border crossings are not equal to the number of unique persons intercepted, as multiple attempts per person can't be resolved and that it depends on the detection probability.

[4]See e.g. `https://www.migrationpolicy.org/pubs/TCMirregularmigration.pdf`

[5]`https://ec.europa.eu/eurostat/statistics-explained/index.php/Asylum_statistics#:~:text=In%202019%2C%20676%20300%20asylum,%2Dto%2Dyear%20since%202015.`

Irregular migration patterns are not stable, as migrants and smugglers continuously search for new ways to enter the European Union and origin country conditions vary or deplete. In recent years, migration patterns shifted between the different routes to Europe [6].

***Prediction of migration patterns***

Frontex aims to operate in an intelligence driven way. This materializes for example in smart detection systems for boats, to be able to intercept refugees crossing the Mediterranean Sea. Next to close to real-time prediction, prediction of such patterns can aide the border police's operations and provide humanitarian relief.

(Disney et al., 2015) give a literature overview of forecasting of migration flows. They distinguish between deterministic and probabilistic approaches. The latter do not take intro account uncertainty of the predictions but rather model different scenarios around a most likely scenario. The probabilistic time series models are based on ARIMA models, which are widely available in statistical software. ARIMA models have a memory in terms of past migration flows; predicted flows in t+1 in a ARIMA(0,1,0) depend on the realize flows in t, a drift, and an error term. However, as this is a model simply extrapolating the past time series, it does not take into account novel information. Those models are often extended using various (demographic) covariates, are based on Bayesian models which work with limited data and take into account expert judgement and population growth projections. The authors compare how several autoregressive models perform on UK immigration and note that errors compound far into the future. Furthermore, they note that different types of migration, such as irregular migration, is much more dependent on events that are poorly predictable. The scope of those models goes up to many decades into the future.

A recent strand of literature aims to improve on those forecasting models on the short run, by employing machine learning procedures using novel data such as databases of news

[6] See, for example: `https://frontex.europa.eu/we-know/migratory-routes/western-balkan-route/`, or more specifically for irregular migration from African irregular migrants to North Africa on page 50: `https://frontex.europa.eu/assets/Publications/Risk_Analysis/Risk_Analysis/Annual_Risk_Analysis_2020.pdf`

events and online search behaviour alongside more traditional data that are known to affect migration (see, e.g. Docquier, Ozden and Peri (2014)). An extreme example of short-run prediction of migration is Ahmed et al. (2016). The authors use lagged migration flows, weather and news data (from the GDELT project) to predict one-day ahead refugee arrival on the Greek islands during the end of 2015 and the beginning of 2016.

Other work considered online search behaviour data to predict migration intentions (Böhme, Gröger and Stöhr, 2020), showing that relative search frequencies as proxied by the Google Trends Index (GTI) have predictive power in a gravity-like regression model in excess of origin-year, origin-destination and origin-destination fixed effects. (Böhme, Gröger and Stöhr, 2020) focus on an economic-theory founded prediction model, and report joint significance of many GTI-based features on both migration flows as well as migration intentions as probed by the Gallup World Poll (GWP). Additionally, several papers suggest the presence of a origin country-dependent lag between migration-related search behaviour and actual migration (Wladyka, 2017; Wanner, 2020).

Using state-of-the-art Machine learning methods, such as (boosted) forest-based regression methods (such as XGBoost (Chen and Guestrin, 2016)) and Artificial Neural Networks (ANNs), Robinson and Dilkina (2017) show that those methods predict both within-USA as well as international migration better than gravity and radiation models common in literature to estimate causal effects on migration flows. The authors use a Common Part of Commuters (CPC) metric to define a loss function to train the model, outperforming commonly used loss functions such as the root mean square error. As Neural Networks usually show less diminishing returns to scale than non-neural network-based methods on more unstructured data, it is expected to overperform tree-based methods only at large amounts of data. The authors find the ANN to perform better than the forest-based XGBoost algorithm in predicting migration flows.

However, as online search behaviour that is indicative of later migration behaviour may have a complex temporal structure, taking into account lagged independent variables may

add predictive power (Wladyka, 2017; Wanner, 2020). Recurrent Neural Networks (RNNs) are able to capture such dynamic effects by modifying ANNs to use the state of the previous time step in the current time step. LSTMs are particularly suitable to model long-run dynamic effects in time series as they overcome the vanishing gradient problem. Examples include speech recognition, handwriting recognition, and polyphonic music modeling (Greff et al., 2017), but also financial time series modelling (Siami-Namini and Namin, 2018).

This is a promising method to improve predictions of international migration, performing very well on one-year ahead bilateral migration flows to OECD countries (Golenvaux et al., n.d.). Golenvaux et al. (n.d.) show that a model including origin, destination, time fixed effects, population and GDP at the origin-year and origin-destination-year level GTI keywords The predictions of Golenvaux et al. (n.d.), trained on 1980-2014, are very well matching actual migration flows in 2015. However, it does not consider how well the model predicts compared to a model that excludes the GTI data.

***Feature importance and interpretation***

A drawback of most prediction methods is a lack of understanding which variables drive the prediction and in which way. Trained models perform highly nonlinear operations to predict on a test dataset. Tree-based methods, such as Random Forest and XGBoost allow for the calculation of a metric based on how often a specific feature is used to split the data to reduce a measure of dissimilarity (usually root mean square error), called variable importance. Using bilateral migration data, Kiossou et al. (2020) employ Partial Dependency Plots (PDPs) to interpret the predictions obtained by a 3-layer deep ANN and find partial dependencies for current stock of migrants from origin i in destination j, distance and GDP in the origin country to affect migration intentions in line with economic theory. Recently, interpretability of machine learning methods caught a lot of attention. LIME and SHAP are two methods able to interpret the output of a machine learning regression or classifier, to assess and interpret predictor behaviour and understand which features are important in specific parts of the dataset (Ribeiro, Singh and Guestrin, 2016). This improves upon the

PDPs as it allows to also consider the whole distribution of marginal effects and can shed light on interactions with other features. Using those methods, we can build up a "fingerprint" of search queries relevant for specific migration flows. (Guo, Lin and Antulov-Fantulin, 2019) suggest a method to interpret the temporal structure of LSTMs.

***This work***

In this proposal, we are combining the monthly Frontex IBC data with GTI data and use LSTMs and interpretable machine learning methods to predict future IBCs on the EU's outer borders. The Frontex data has one large advantage: the data is on a monthly level, allowing for using the temporally fine GTI data (in principle, this data can be retrieved at the day-level). Moreover, smart phone usage and the ability to query information and to communicate online is shown to be very important for irregular migrants in their journey towards Europe (Zijlstra and Liempt, 2017). Due to rapid recent expansion of internet coverage in Africa (Manacorda and Tesei, 2020), many potential irregular migrants have access to the internet. Worldwide, around 25% of world population used the internet in 2009, whereas this figure exceeds 50% nowadays [7]. Partially because of a heavy reliance on mobile internet, Google has a dominant market share in Africa in 2021 and a very large market share in most Middle East and South Asian countries in scope [8]. Hence, GTI provides meaningful information of search behaviour within the countries of interest. However, irregular migration patterns are by definition smaller than bilateral yearly migration flows, so a large part of the origin-country internet traffic is not performed by potential migrants, potentially masking the relevant search behaviour. Similarly, the prediction literature acknowledges the difficulties of predicting irregular migration (Disney et al., 2015).

A limitation of using GTI is comparisons of key word searches between countries and languages. Given a set of key words $W_{EN}$ in English, we can't easily find an equivalent $W_{TR}$ in a different language. For example, a (literal) machine translation of a keyword to another

[7] `https://data.worldbank.org/indicator/IT.NET.USER.ZS`

[8] For regional and country-level market shares in the search engine market, see: https://gs.statcounter.com/search-engine-market-share

language leads to a vastly different number of hits because of different non-literal wording. For a given search term, very little origin-destination pairs show up nonzero as we may query in a non-native language. This weakness of (Böhme, Gröger and Stöhr, 2020) is amplified by the limitations of GTI to find relative frequencies of rare keywords (see the next section). Nevertheless, as one can include country fixed effects, a neural network model can learn that some key words for some countries are not indicative of migration intentions.

Understanding what terms are most likely to predict subsequent migration is very relevant to policy. As an example, Search Engine Optimization (SEO) on important key words may be used to push encouraging or discouraging information and thus affect the development of migration intentions by potential destination countries [9].

In the following, we assume the reader to have some basic understanding of neural networks. For some more background information, we refer the reader to Mullainathan and Spiess (2017) and Athey and Imbens (2019) for an overview of such methods.

# 2 Data

## 2.1 Realized migration flows to OECD countries

We obtained data of IBCs from Frontex. This data comprises monthly time-series between January 2009 and November 2020 for 496 country of origin-route pairs. The routes are: Western Africa (Canary islands), Western-, Central-, and Eastern Meditteranean, Western Balkan, Eastern borders, Black Sea and the Circular route from Albania to Greece. Note that the country-routes pairs included are those that have at least one detected IBC in the period of study. We sum IBCs attempted over land and over sea for each country-route pair, and exclude origin-countries that are 1) microstates (we define those as having a population of less than 500,000) and 2) not located in Africa, Middle East, Europe, or South Asia.

[9] For an example of such an information campaign, see `https://core.ac.uk/download/pdf/42579387.pdf`

## 2.2 GTI

To understand the GTI index, the underlying data collection and the normalization procedure should be properly understood (Steegmans, 2019). The GTI represents a normalized time series of relative google search frequency, sampling an unknown percentage of queries. It is indexed on an integer scale of 0 to 100, where 100 is normalized to the highest relative search frequency of the requested data. This means that the resolution of the data is limited to 1% of the peak value. This is a particular problem if the time series data is very variable, which does not allow to resolve small differences in times with relatively low search frequencies. A possible workaround to this is described in Appendix B.

For example, if we are interested in the relative frequency of both search query A and query B, we need to query those simultaneously and the data is scaled on the highest value occurring in the series of either A or B[10].

As a proof of principle, we query GTI for all origin countries in our data set for a set of keyword listed in Appendix A. As we want a big data approach, in this project we later want to take into account different keywords capturing different search intentions. We lag the GTI data by one month, as we aim to predict IBCs based on past search behaviour and don't want to include simultaneously available information reflected in the GTI data.

In our models, we include fixed effects at various levels to obtain a baseline prediction. Those levels are: 1) Country of origin, 2) Route, 3) Month of the year, 4) Year. This allows us to control for determinants of migration at several levels. In this proof of principle, we include the GTI of keywords that vary only at the country of origin-time level.

[10]Google Trends has a query limit, documented to be around 1500 queries within a 12 hour time frame on a single IP address. After that, one can perform a single query per 60 seconds. To circumvent this, our code flags hitting the IP address limit, after which the user can change his IP address. This can be done by either using a VPN, an IP address crawler, or manually changing the IP address by e.g. re-initializing a mobile cellular connection after the limit is reached. This allows to scale up GTI to include many keywords.

### 2.3 Data normalization

We normalize all data used in the model to a (0,1) scale using a min-max-scaler, as this improves the behaviour of a neural network. Model performance can deteriorate if data is not normalized, as features of different scales may be treated unequally, leading some weights to be trained sub-optimally.

### 2.4 Data partitioning

To train and evaluate a predictive method, we split the training and the test data into two parts. We can proceed in three distinct ways: 1) split the data set randomly 2) split the data set on observations 3) split the data set chronologically. In this proposal, we exclusively focus on 3) as it is most relevant to policy, because we aim to predict future flows for all origin-route pairs. If not all origin-route pairs would have been observed, 2) would be a relevant to consider how the model generalizes to new country-route pairs.

In our model, we only do one-month-ahead testing. However, as a single month may be inherently noisy (e.g. because of the introduction of new Frontex missions), the prediction error remains considerable. Moreover, migration patterns for some months of the year may be more predictable then others. Ultimately, longer prediction intervals may be more informative and easier to predict.

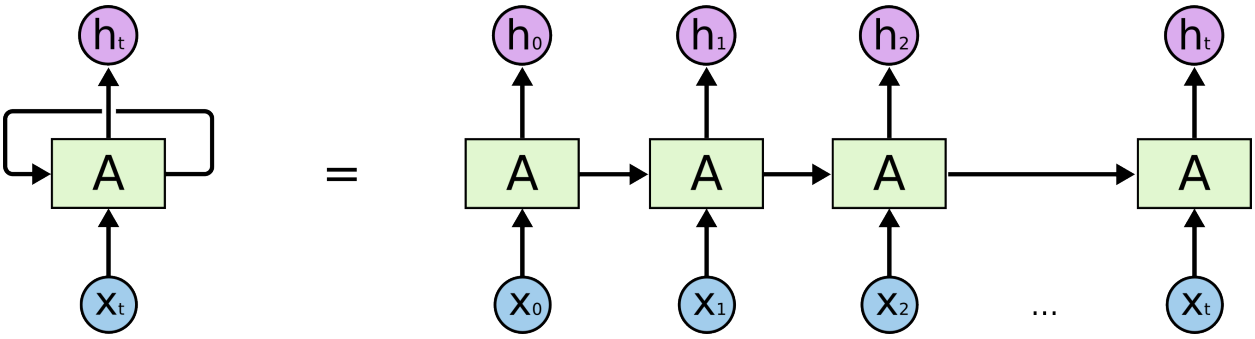
## 3 LSTMs

LSTMs are a type of Recurrent Neural Networks (RNNs), which are a type of neural networks that are able to exploit time series data. LSTMs show superior performance on tasks where longer lags are important (Greff et al., 2017), as it counteracts the vanishing gradient problem in the back-propagation step of the optimizer (stochastic gradient descent or varieties thereof). This is particularly appealing to our case, as the process from migration intention (and related search behaviour) to an illegal border crossing may take months or

even years.

Figure 1 depicts a basic Recurrent Neural Network layer, where $x_t$ are the input vectors and $h_t$ the output vectors. The block A contains the Neural Network; it is the same in every time step. The network saves the state of block A and passes it on to the next time step: the network is trained in a recurrent way. However, as all the blocks contain the same neural network(s) with the same weights, gradients are either small and in the updating (backpropagating) step the compounded weights update becomes too small, or gradients are large and the compounded update becomes very large. This limits the learning to only a few time steps. In other words, we can't use meaningful information more than a few time periods ago.

Figure 1: Schematic of a folded RNN (from Colah (2015))



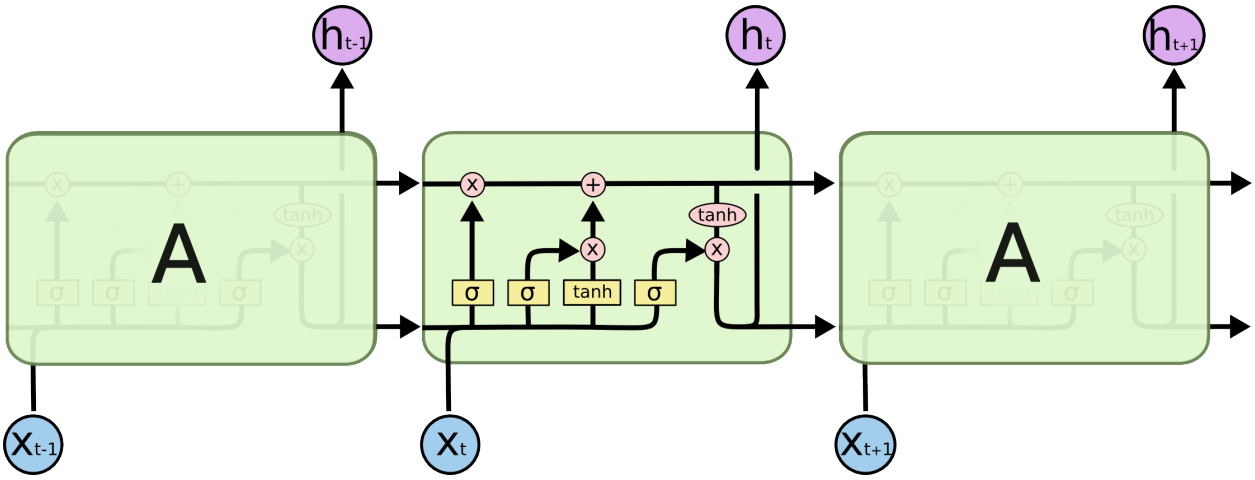

Hochreiter and Schmidhuber (1997) found a solution to this problem by introducing the Long Short Term Memory network (LSTM). Figure 2 shows the rather complex LSTM network geometry, which connects two temporally adjacent cells by two instead of one lead: one contains the lagged predicted output, whereas the other contains the state obtained from the previous cells. Here, $\sigma$ denotes a sigmoid activation function mapping to (0,1) and *tanh* (the hyperbolic tangent) which maps to (-1,1). These functions have desirable properties for learning, as their gradients are easy to compute and are monotonic. Those functions are applied to a weighted transformation of its inputs: those weights are updated when training the network.

The upper lead is called the cell state C_t and is modified by a multiplicative gate that

is coined the forget gate, as it decides how much of the previous state to save. The next gate decides what information to add to the cell state, based on the lagged output value and the inputs. Again, this goes through a sigmoid function, but also to a hyperbolic tangent function that scales its output to (-1,1) and the new state is saved and is not strictly positive because of the use of the tanh. Using the new state, the input data and the predicted lagged dependent variable, the output $h_t$ is predicted. The lower lead is the predicted lagged output $h_{t-1}$.

Because of the architecture with a separate cell state, the gradients in the backpropagating step do not decrease exponentially, as is the case for regular RNNs. Note that the dimensionality of the state and output vectors is $m$, the number of nodes in the LSTM layer. Thus, the output $h_t$ has the same dimension, which can be connected to a subsequent layer (such as a dense lazer connected to a loss function).

Figure 2: Schematic of a LSTM cell (from Colah (2015))




A drawback of RNNs is, is that because they are defined in a sequential way, it is not possible to easily parallelize the training step. This makes LSTMs to scale less good as conventional (deep) neural networks for very large amounts of data.

### 3.0.1 Our model geometry

Our implementation of LSTMs follows (Golenvaux et al., n.d.), which uses a one-layer LSTM. Usually, using more complex (deeper) models are able to uncover more complex relationships between the variables of unstructured data. As we work with relatively structured data, namely the relative proportion of search queries of online search traffic, we deem a single layer sufficient. Moreover, models with multiple layers are harder to regularize and also have a larger hyperparameter space to optimize. Although some attempts for a two-layer stacked LSTM were made, this didn't prove to improve prediction.

### 3.0.2 Regularization in LSTMs

As we want our model to generalize to new, unseen data, we have to prevent overfitting the training data set. There are various possibilities to prevent overfitting. However, regularization has to be tuned appropriately. When regularization is too strong, the model loses the ability to learn the data or it takes many epochs to learn.

First, by keeping the model simple in terms of depth and number of nodes per layer there is little possibility to exactly fit the data, but makes the model less flexible. A powerful method is to drop out part of the nodes in each batch. In this so-called drop-out a random share of nodes are not updated, such that those notes don't exactly learn to fit the training data. The tunable parameter is the share of input nodes (so before the LSTM layer) dropped out of the updating step. We use a drop out rate of 20%.

Moreover, we can employ regularization on the weights in the intermediate nodes by adding ridge- and/or lasso-regularization terms to the nonlinear activation functions in the nodes. We use a lasso (L1) regularization on all weights in the model, suppressing model complexity and penalizing large weights. Other possible methods are e.g. adding noise between the layers of the model and using recurrent dropout, which removes at random some of the instances in the temporal dimension of the LSTM layer.

## 3.1 Loss functions

Any regression-based neural network requires an objective function that represents a loss that is to be minimized in order to perform stochastic gradient descent by backpropagating the gradient. A simple measure is the Root Mean Square Error (RMSE), which is defined as sum of squared differences between predicted and realized data. In this work, we use the Common Part of Commuters (CPC) function that is a measure of similarity between two two-dimensional matrices. The CPC compares a realized matrix T and an estimated matrix $\hat{T}$ with rows i, indexing origins and columns j, indexing destinations. We define it as follows (Robinson and Dilkina, 2017):

$$\boldsymbol{CPC}(\boldsymbol{T}, \hat{\boldsymbol{T}}) = \frac{\mathbf{2} \sum_{i,j=1}^{n} \mathbf{min}(\boldsymbol{T_{ij}}, \hat{\boldsymbol{T}}_{ij})}{\sum_{i,j=1}^{n} \boldsymbol{T_{ij}} + \sum_{i,j=1}^{n} \hat{\boldsymbol{T}}_{ij}} \tag{1}$$

We define the loss function associated with the CPC metric as $1 - CPC$. The CPC loss functions are found to behave better in sparse data with a lot of 0 values (Robinson and Dilkina, 2017).

## 3.2 Hyperparameters

LSTMs, as many other machine learning methods, have several model parameters that can be and need to be tuned to improve performance. The parameters in the case of an LSTMs are:

- Model geometry: the number (depth) and types of layers

  One layer.

- Model geometry: the number of nodes per layer

  70 nodes.

- Model optimization: the optimizer and its learning rate

  We use the adam optimizer, which is an extension of stochastic gradient descent that

works particularly well by modifying learning rates based on the local gradient (Kingma and Ba, 2014). We use a learning rate of 0.005.

- Model optimization: the batch size

  Based on a batch, the weights are updated. The larger the batch, the less computational effort is needed to run through the full training sample once (called an Epoch). The smaller the batch, the more often weights update. However, small batch sizes may lead to noisy. We define a single batch to exist out of one time series for one country-route pair.

- Model optimization: the loss function

  We use the CPC loss function described above.

- Model regularization: the dropout rate

  A regular dropout of 0.2

Although the parameters mentioned above work reasonably well, a hyperparameter search should be performed, using e.g. a grid search. However, this is computationally costly. We implement the one-layer LSTM in the python package Keras, with Tensorflow as a backend.

# 4 Preliminary Results

In order to examine whether Google Trends Indices are able to improve predictions we consider two models: a first model with only fixed effects for month, country and route, and a second model comprising the first model plus the Google Trend Indices for the keywords listed in Appendix A. We plot two different metrics of similarity of both models that we run for 70 epochs. Figure 3 shows the learning curves in terms of RMSE and CPC of the model without GTI data, whereas Figure 4 shows the learning curves for the model with

GTI data[11]. For both models, we observe the test RMS error to be lower than the training error, which is due to our test set consisting of only a single month, whereas the training set includes all months. The seasonal variation over a year inflates the training error. The learning curves are relatively noisy, which may because there may be many local minima in the parameter space of possible weights: the GTI data for different keywords is strongly correlated as well as (linear) combinations of the fixed effects. Hence, distinctly different weights may lead to a similar prediction, leading the algorithm not to converge to a single minimum, but to move around many.

The root mean squared error and CPC measure show distinct behaviour in their learning curves. Interestingly, the RMSE of the test set keeps decreasing, whereas the CPC (that is used as a loss function) already flattened out after 10 epochs.

In figure 5, we show that on the test set the CPC is lower when including GTI data by around x% (and around 20% in terms of root mean squared error). Figure 6 shows a log-log scatter plot of predicted versus observed IBCs using the model with GTI data. We observe that we reasonably well predict the

# 5 Discussion and Outlook

In this proposal we show a proof of principle that Google Trends Indices can be used not only to predict international migration flows, but also irregular migration taking the form of illegal border crossings on Europe's outer border. However, two general remarks should be made to assess the usefulness of a predictive model.

First of all, we haven't yet compared the model including GTIs to a model of full information at the moment the GTI has been measured. Any predictive method that uses data from $t - t'$ and before to predict in period $t$ using new information (in our case GTI data) from $t - t'$ should be able to improve prediction given all other available data in $t - t'$. If this

[11]Note that those curves are not unique, as the stochastic nature of the optimizer, dropout and initialization of weights is different in every instance. However, using many repetitions, the GTI model performs better.

is not the case, the new information is already conveyed in the other available data. In this work, we have set t' to one month. However, it may very well be that given the knowledge of IBCs in $t - t'$ (the $t'$ lagged dependent variable) we can't improve the prediction using $t - t'$. Therefore, in a final model it is important to vary $t'$. If for some $t'$ in a model that includes all lagged dependent variables up to $t'$ as independent variables we can enhance prediction, we have a useful model.

Secondly, the results lack clear interpretation on how and why relative search behaviour affects. Although the results may be useful for e.g. allocating border police capacity over time and location, it does not shed light on the irregular migration decision and associated behaviour by the potential migrant population in origin counties

We believe additions on three fronts would improve the predictive model and the relevance of those results:

**Technical**

- Fully implementing the GTAB package reduces measurement error in Google Trends indices by anchorbanking the full distribution of search intensities (West, 2020)

- After inclusion of many different (groups of) GTI keywords, we want to select the most important features by using interpretable machine learning methods, such as SHAP and LIME, to be able to consider the marginal effects of specific search queries (Ribeiro, Singh and Guestrin, 2016).

- Furthermore, we are particularly interested in the timing of search behaviour: Recent modifications of LSTMs allow interpretation of which time lags are important for which features (Guo, Lin and Antulov-Fantulin, 2019).

- An extensive hyperparameter search on a High-Perfomance Computing (HPC) facility to find the strongest model.

**Methodological**

- Inclusion of many more keywords at the origin-month level, relating to push factors of migration.

- Categorization of sets of keywords that proxy for different motives for migration, in order to shed light on the propensity to migrate for different motives.

- Making use of Google Trend categories (such as "Immigration policy and border issues", "job listings", or "Politics") in combination with the destination country or geographic entity names on the route.

- Making use of "control queries" to control for interest on the destination country level that is unrelated to migration intention, such as the interest in sports events. This is closely related to the use of anchorbanking as explained in Appendix B

- Including GTI data for search interest of other migration destinations and associated routes.

- We can extend the model to include what share of population has (mobile and broadband) internet access. As the GTI are relative search frequencies, information about the absolute number of internet users adds predictive power.

**Policy relevant**

- Going beyond improving predictions, knowledge about how search behaviour shapes irregular migration is very valuable. Using interpretable machine learning methods, we can indicate which search queries at what time impact irregular migration patterns most. This is a useful addition to the literature that deals with the determinants of irregular migration as well as useful for policy makers that want to steer migration decisions and possibly alleviate conditions that lead people to undertake a dangerous travel.

# Appendix A: Keywords

## A1: Current Keywords

The keywords we use in the proof of principle model are (based on Böhme, Gröger and Stöhr (2020)):

- legalization
- passport
- visa
- quota
- waiver
- immigrant
- immigrate
- immigration
- arrival
- emigrant
- emigrate
- emigration
- required documents
- migrant
- migrate
- migration

### A2: Categories of keywords to use

The keywords we wish to include can be grouped into three categories:

- GTI origin level: we record the google trend index of origin-country
- GTI destination in category: we source the name of potential destination countries within a google trends category we deem relevant. For example, for the Western Mediterranean route, we query "Spain" in English, Hausa, and Igbo in multiple categories, such as the category "jobs".
- GTI transit: we retrieve the GTI for keywords that include queries for countries that are transit countries, such as Morocco for the Western Mediterranean route.

## Appendix B: Achorbanking

West (2020) suggested a solution to two problems of Google Trends. First of all, the GTI normalizes all simultaneous queries to the most popular query (GTI=100) of that set of queries, obfuscating the lesser common queries' temporal behaviour, as popularity is an integer in [0-100]. Secondly, we can't query more than 5 keywords at once. Thus, an accurate comparison of any two search terms requires a chain of overlapping search queries along the popularity scale. The technical implementation of West (2020) builds up a so-called anchorbank of such queries defined for a specific region and time-period, to cover the whole popularity scale. After construction of the anchorbank, we can always find a similarly popular saved query for a new query in the same geographic and temporal scope. West (2020) supplied a python-package G-TAB, which allows us to improve upon the existing model.

## References

**Ahmed, Mohammed N., Gianni Barlacchi, Stefano Braghin, Francesco Calabrese, Michele Ferretti, Vincent Lonij, Rahul Nair, Rana Novack, Jurij**

**Paraszczak, and Andeep S. Toor.** 2016. "A multi-scale approach to data-driven mass migration analysis." *CEUR Workshop Proceedings*, 1831: 1–17.

**Athey, Susan, and Guido W Imbens.** 2019. "Machine learning methods that economists should know about." *Annual Review of Economics*, 11: 685–725.

**Böhme, Marcus H, André Gröger, and Tobias Stöhr.** 2020. "Searching for a better life: Predicting international migration with online search keywords." *Journal of Development Economics*, 142: 102347.

**Chen, Tianqi, and Carlos Guestrin.** 2016. "Xgboost: A scalable tree boosting system." 785–794.

**Colah, J.** 2015. "Understanding LSTM Networks."

**Disney, George, Arkadiusz Wiśniowski, Jonathan J Forster, Peter W F Smith, and Jakub Bijak.** 2015. "Evaluation of existing migration forecasting methods and models." *Report for the Migration Advisory Committee: Commissioned research. ESRC Centre for Population Change, University of Southampton.*

**Docquier, Frédéric, Çağlar Ozden, and Giovanni Peri.** 2014. "The Labour Market Effects of Immigration and Emigration in OECD Countries." *The Economic Journal*, 124(579): 1106–1145.

**Golenvaux, Nicolas, Pablo Gonzalez Alvarez, Harold Silvère Kiossou, and Pierre Schaus.** n.d.. "An LSTM approach to Forecast Migration using Google Trends."

**Greff, K., R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber.** 2017. "LSTM: A Search Space Odyssey." *IEEE Transactions on Neural Networks and Learning Systems*, 28(10): 2222–2232.

**Guo, Tian, Tao Lin, and Nino Antulov-Fantulin.** 2019. "Exploring interpretable LSTM neural networks over multi-variable data." 2494–2504, PMLR.

**Hochreiter, Sepp, and Jürgen Schmidhuber.** 1997. "Long short-term memory." *Neural computation*, 9(8): 1735–1780.

**Kingma, Diederik P, and Jimmy Ba.** 2014. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980.*

**Kiossou, Harold Silvère, Yannik Schenk, Frédéric Docquier, Vinasetan Ratheil Houndji, Siegfried Nijssen, and Pierre Schaus.** 2020. "Using an interpretable Machine Learning approach to study the drivers of International Migration."

**Manacorda, Marco, and Andrea Tesei.** 2020. "Liberation Technology: Mobile Phones and Political Mobilization in Africa." *Econometrica*, 88(2): 533–567.

**Mullainathan, Sendhil, and Jann Spiess.** 2017. "Machine learning: An applied econometric approach." *Journal of Economic Perspectives*, 31(2): 87–106.

**Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin.** 2016. ""Why should i trust you?" Explaining the predictions of any classifier." *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-August-2016: 1135–1144.

**Robinson, Caleb, and Bistra Dilkina.** 2017. "A Machine Learning Approach to Modeling Human Migration." *arXiv.*

**Siami-Namini, Sima, and Akbar Siami Namin.** 2018. "Forecasting Economics and Financial Time Series: ARIMA vs. LSTM."

**Steegmans, JWAM.** 2019. "The Pearls and Perils of Google Trends: A Housing Market Application." *USE Working Paper series*, 19(11).

**Wanner, Philippe.** 2020. "How well can we estimate immigration trends using Google data?" *Quality and Quantity*, 0123456789.

**West, Robert.** 2020. "Calibration of Google Trends Time Series." *International Conference on Information and Knowledge Management, Proceedings*, 2257–2260.

**Wladyka, Dawid K.** 2017. "Queries to google search as predictors of migration flows from Latin America to Spain." *Journal of Population and Social Studies*, 25(4): 312–327.

**Zijlstra, Judith, and Ilse Van Liempt.** 2017. "Smart(phone) travelling: understanding the use and impact of mobile technology on irregular migration journeys." *International Journal of Migration and Border Studies*, 3(2-3): 174–191.

Figure 3: Learning curves in terms of RMSE and CPC loss for the model without GTI

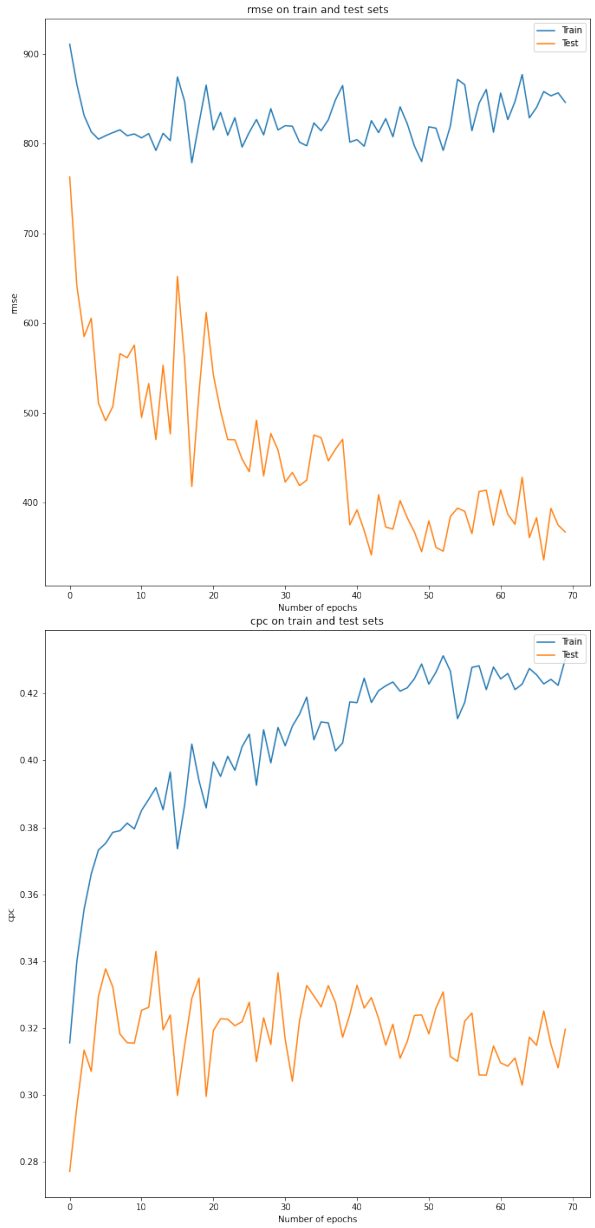


Figure 4: Learning curves in terms of for model with and without GTI

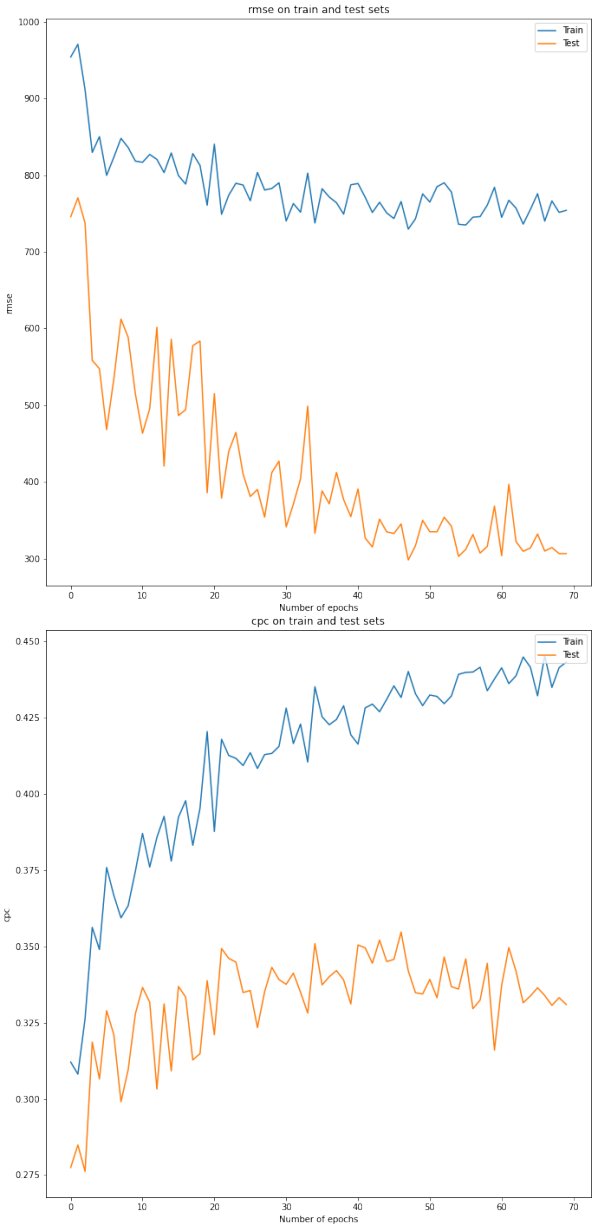


Figure 5: Comparing with and without GTI

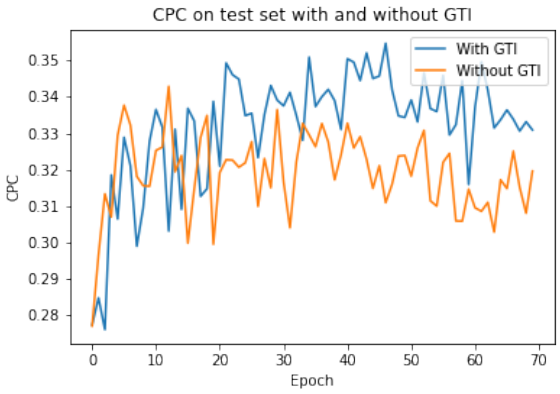


Figure 6: Predicted versus actual IBCs on the test data set in the model