December 31, 2023

SAN Final Assignment - Report

David Čech
cechdav5@fel.cvut.cz

Viktor Korladinov
korlavik@fel.cvut.cz

Diana Korladinova
korladia@fel.cvut.cz

Tomáš Mlynář
mlynatom@fel.cvut.cz

Contents

List of Figures

List of Listings

List of Tables

1. Introduction

In the past two years, a significant number of Ukrainian refugees have taken shelter in the Czech Republic due to the ongoing war in their country. Since this phenomenon has had a notable impact on the Czech Republic's economy, among other things, the main goal of this project is to assess whether the growing number of incoming evacuees has affected the unemployment rate in Czechia and in what way. The idea was inspired by the observations of a Labor Office employee stating that the number of unemployed women has been steadily increasing, and their hypothesis is that one of the main causes is the influx of Ukrainian workers. As part of our work, we release the collected dataset and code on our GitHub page[1].

2. Problem Definition

The primary objective of this project is to gauge whether the growing number of incoming Ukrainian refugees directly affects the growing unemployment rate of women in the Czech Republic or whether it is a result of the country's changing economic situation.

3. Gathering Information

Because no complete dataset was available, we created a novel dataset by combining publicly available data into a complex dataset. For this project, we are mainly interested in three types of data: the unemployment rate in recent years, the number of Ukrainian refugees, and economic statistics (in case the increasing unemployment rate is only affected by the current economy). With that in mind, we chose the majority of our datasets from the Ministry of the Interior of the Czech Republic[2], the Ministry of Labour and Social Affairs of the Czech Republic[3][4], and the Czech Statistical Office as our data sources. Because the data were provided in different formats and often in a bad form (not easily accessible for our scripts), the work was demanding and tedious. We hope that in the future, the state will publish the data in computer-readable formats, fulfilling the standards.

3.1. Dataset Description

Given the objective of the task, the dataset is comprised of the following columns where the refugee-related columns are collected only for the duration of the Russian-Ukrainian war (03.2022-09.2023), and the other predictors are collected from **01.2009** to **09.2023**[5]. If some data were provided only for years or quarters, we have performed simple imputation, expecting that in the given period, the data is constant. The actual predictors are as described below:

- Temporal data: *month*, *year*
  - Because we are working with time series, we will treat these as ordinal data (month as the number of months from 1 to 12 and year as the number of years).
- Geographical: *kraj*
  - We will work with statistics for the entire Czech Republic specified for regions NUTS3 (kraje). We will encode them according to the reference sheet of the Czech Statistical Office[6]. During the model development, this column will be treated as categorical.
- Unemployment monthly statistics (Dataset Link)
  - *uchazeciOZamestnaniUoZ* - the total number of female job applicants by the end of the month
  - *noveHlaseniUchazeci* - the number of new job applicants in the month
  - *noveHlasenaAUvolnenaVPM* - the number of new jobs in the month
  - *obsazenaAZrusenaVPM* - the number of cancelled job places

[1] https://github.com/DianaKorladinova/SAN_Unemployment_Refugees
[2] https://www.mvcr.cz/clanek/statistika-v-souvislosti-s-valkou-na-ukrajine-archiv.aspx
[3] https://data.mpsv.cz/web/data/otevrena-data4
[4] https://data.mpsv.cz/web/data/otevrena-data16
[5] These time periods are motivated by the data availability from Czech government and its agencies.

- *absolventiSkolAMladistvi* - the number of applicants from schools
- *uchazeciOZamestnaniUoZZeny* - our dependent variable, the number of female job applicants by the end of the month
- *uchazeciOZamestnaniUoZMuzi* - the total number of male job applicants by the end of the month
- *uchazeciOZamestnaniUoZMuzi_ratio* - the ratio of male job applicants with *uchazeciOZamestnaniUoZ* by the end of the month
- *uchazeciOZamestnaniUoZZeny_ratio* - the ratio of female job applicants with *uchazeciOZamestnaniUoZ* by the end of the month

- Ukrainian Refugees Statistics (Dataset Link)
  - *celkem* - the number of Ukrainian refugees that come in the month to the kraj (of any age) (additionally *celkem_w?* are distributed lags for window sizes from 2 to 19 = meaning how many refugees came in the last **?** months)
  - *m_do_65* - the number of male Ukrainian refugees between 18 and 65 years (additionally *m_do_65_w?* are distributed lags for window sizes from 2 to 19 = meaning how many men refugees came in the last **?** months)
  - *z_do_65* - the number of female Ukrainian refugees between 18 and 65 years (additionally *z_do_65_w?* are distributed lags for window sizes from 2 to 19 = meaning how many women refugees came in the last **?** months)
  - *m_do_65_ratio* - the ratio of male Ukrainian refugees (with celkem) between 18 and 65 years (additionally *m_do_65_w?* are distributed lags for window sizes from 2 to 19 = meaning how many male refugees came in the last **?** months)
  - *z_do_65_ratio* - ratio of female Ukrainian refugees between 18 and 65 years (additionally *z_do_65_w?* are distributed lags for window sizes from 2 to 19 = meaning how many women refugees came in the last **?** months)
- Monthly Development of Difference in Inflation w.r.t Last Year: (Dataset Link)
  - *monthly_inflation_rate_wrt_last_year* - inflation as defined by the Czech Statistical Office (2nd table in the link)
- Criminality Rate (Dataset Link) -*break_in_thefts* - the number of break-in thefts in the Czech Republic in the given time period *general_thefts* - the number of general thefts in a given time period
- Salaries (Dataset Link)
  - *avg_monthly_salary* - average salaries per quarter for every region w.r.t to the number of employed people (2nd column in the data source) in Kč
- Minimum Wage (Dataset Link)
  - *monthly_min_wage* - the minimum wage in the given month in Kč
- Energy
  - Cost of Electricity (Dataset Link) and (Dataset Link)
    - *avg_energy_price* - average electricity price in the given month in USD/1Mwh
  - Cost of Natural Gas (Dataset Link) and (Dataset Link)
    - *avg_natural_gas_price* - average natural gas price in the given month in USD/1MMBtu
  - Cost of 1l of Gasoline (Dataset Link) and (Dataset Link)
    - *avg_gasoline_price* - average gasoline price in the given month in USD/1l
- Zahraniční obchod (obchodní bilance) - (Dataset Link)
  - *bilance* - the balance of import and export of the Czech Republic
- REER (Definition) (Dataset Link)
  - *reer* - Real effective exchange rate of the Czech Crown.

Overall, our dataset consists of 117 predictors (columns) and 2478 data points (rows). Below in Listing 1, we provide a single sample from September 2023 in JSON format.

[6] https://www.czso.cz/documents/10180/20536384/13-72390704.pdf

```
{
  "month": 9, "year": 2023, "kraj": "ZLK", "general_thefts": 67.0,
  "break_in_thefts": 57.0, "avg_monthly_salary": 38596.0, "reer": 115.41,
  "celkem": 162.0, "m_do_65": 70.0, "z_do_65": 65.0,
 "celkem_w2": 314.0, "celkem_w3": 628.0, "celkem_w4": 895.0, ...,"celkem_w19": 11226.0,
  "m_do_65_w2": 126.0, "m_do_65_w3": 226.0,..., "m_do_65_w19": 2955.0,
  "z_do_65_w2": 120.0, "z_do_65_w3": 276.0,...,"z_do_65_w19": 4907.0,
  "monthly_min_wage": 17300, "monthly_inflation_rate_wrt_last_year": 6.9,
  "m_do_65_ratio": 0.4320987654320987, "z_do_65_ratio": 0.4012345679012346,
  "m_do_65_w2_ratio": 0.4012738853503185,..., "m_do_65_w19_ratio": 0.2632282202030999,
  "z_do_65_w2_ratio": 0.3821656050955414,..., "z_do_65_w19_ratio": 0.4371102797078211,
  "bilance": 11.874, "avg_energy_price": 138.98, "avg_gasoline_price": 1.7623,
  "avg_natural_gas_price": 2.7372, "uchazeciOZamestnaniUoZ": 10522,
  "uchazeciOZamestnaniUoZZeny": 5666, "uchazeciOZamestnaniUoZMuzi": 4856,
  "noveHlaseniUchazeci": 2341, "noveHlasenaAUvolnenaVPM": 866,
  "obsazenaAZrusenaVPM": 911, "absolventiSkolAMladistvi": 811,
  "uchazeciOZamestnaniUoZMuzi_ratio": 0.4615092187796997,
  "uchazeciOZamestnaniUoZZeny_ratio": 0.5384907812203004
}
```


Listing 1: Dataset sample in JSON format. The three dots (...) represent an interval of windows of the predictor

3.2. License and Mentions

In compliance with the license conditions of ČSÚ, we are disclosing that all of the data sourced from them has been acquired via their web interface; additionally, all of the data has been used as-is, without any modifications that could result in a different interpretation[7]. Another source of our information is the system ARAD of the Czech National Bank (ČNB), and the only data collected from there is the REER predictor. The manipulation with it was also done in accordance with the conditions[8], and no modifications were made. The other data are publicly available without any license conditions that should be mentioned (mainly, they are provided under the auspices of the organization Open Data[9]).

4. Risks and Limitations

We cannot account for every risk, but we did our best to outline possible problems and limitations to our project.

- Some of the datasets we want to include (such as HDP) are not publicly available.
- There is a significant risk of confounders - for example, the unemployment rate is influenced by a vast range of predictors, and we cannot possibly address all of them.
- The war spans over a short time interval, so the data related to it is minimal. Because of that:
  - The error distribution cannot be observed.
  - There is not much information to train models on.
- As always, model assumptions may not reflect reality.

[7] https://www.czso.cz/csu/czso/podminky_pro_vyuzivani_a_dalsi_zverejnovani_statistickych_udaju_csu
[8] https://www.cnb.cz/arad/#/cs/documentation
[9] https://opendata.gov.cz/informace:start

5. Preliminary Analysis

Preliminary analysis plays an integral part in every statistics-oriented project. To be able to select an appropriate model and train it on the collected datasets, we need to analyze and understand the data they contain. A powerful way to dissect and examine our features is by visualizing them - the data we are working with is temporal; thus, it pays to search for recurring patterns or seasonal spikes. It is also essential to explore the correlation between the different factors we've decided to include. Our justification for this extra work is that we must be aware of all these relations in order to be able to avoid making statistical errors.

There are many ways to visualize time-related data. We have selected the methods relevant to our case and outlined the results and the impact of the knowledge gained from them below. Due to document volume constraints, we cannot display the entire analysis we conducted. Thus, we will provide some examples in each category and leave the detailed report in our GitHub repository.

5.1. Time Plot and Pattern Analysis

As mentioned above, we are working with time series. The goal of the current analysis is to observe trends, cycles, and anomalies and attempt to explain them via the socioeconomic situation at the time of record. It is vital to strike a balance - finding a pattern would be very beneficial to us as it would mean we can predict future values with high accuracy based on previous ones; there is, however, the risk of making a type I error – the false identification of pattern that does not exist in reality. To minimize the probability of this happening, we have ensured that the time interval is big enough and encompasses enough data to focus on long-term trends, which tend to be much more stable than short-term changes.

5.1.1. Average wage

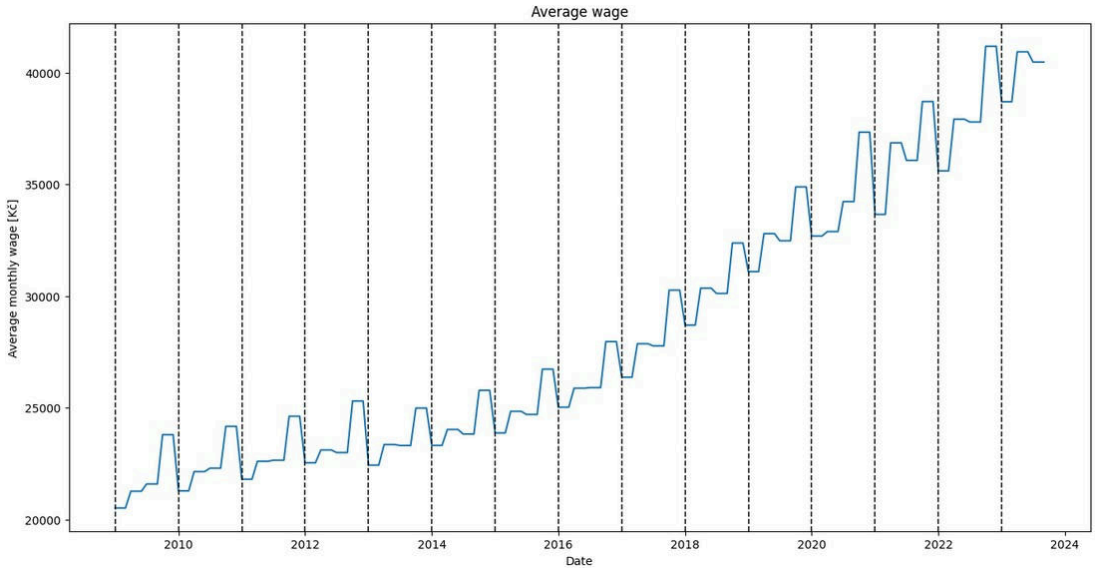

Figure 1: Time plot of the average wage: *avg_monthly_salary*

A simple example of an obvious seasonal pattern with an apparent upward trend. Every year, the average wage increases, which makes sense from an economic point of view.

5.1.2. Jobseekers

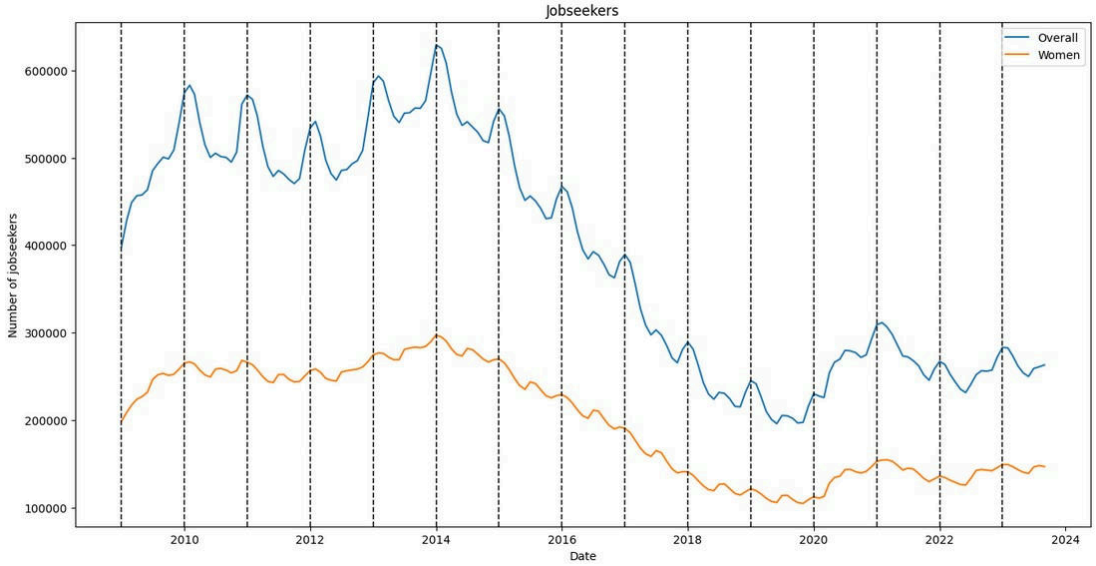

Figure 2: Time plot of jobseekers: *uchazeciOZamestnaniUoZ* and *uchazeciOZamestnaniUoZZeny*.

Strong seasonal dependency can be observed in Figure 2, with a downward trend after 2014; the trend was interrupted in 2020, which can be attributed to the start of the global pandemic, COVID-19. Another hike is present in the recent data - 2022 saw an increase. That may be due to the heightened refugee immigration stemming from the Russian-Ukrainian war.

5.1.3. Newly announced jobseekers

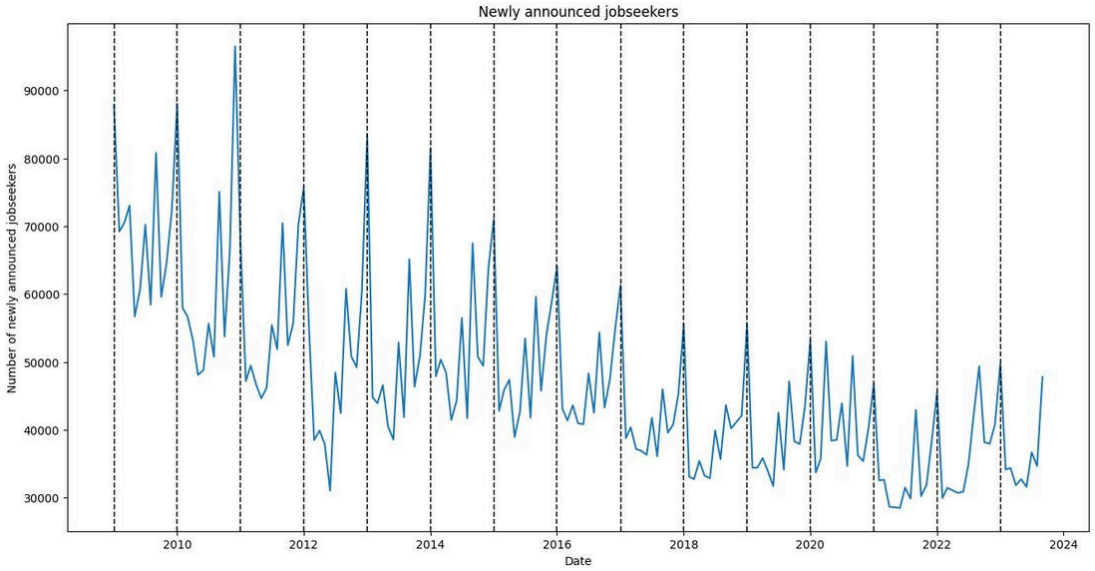

Figure 3: Time plot of newly announced jobseekers: *noveHlaseniUchazeci*

Similar to the previous graph, in Figure 3 there is both a seasonal pattern and a downward pattern. The data here increases our confidence in our hypothesized cause of the recent upward tendency - 2020 had many waves of newly announced jobseekers, which corresponds to the true events, where many businesses closed down throughout the year, letting go all of their employees. The year 2022 displays two big waves in its second half, which is around the time the waves of refugees arrived.

5.1.4. Ukranian Refugees

The plot in Figure 4 shows a major influx of refugees at the start of 2022, which supports our previous claims. Due to the reasons for migration, the vast majority of migrants are women. The inverse spike is

caused by the expiration of the temporary protection visas granted by the government - their validity is year-long and ends in March unless extended for another year (Source). The spike is caused by the refugees who did not extend their visas and migrated instead.

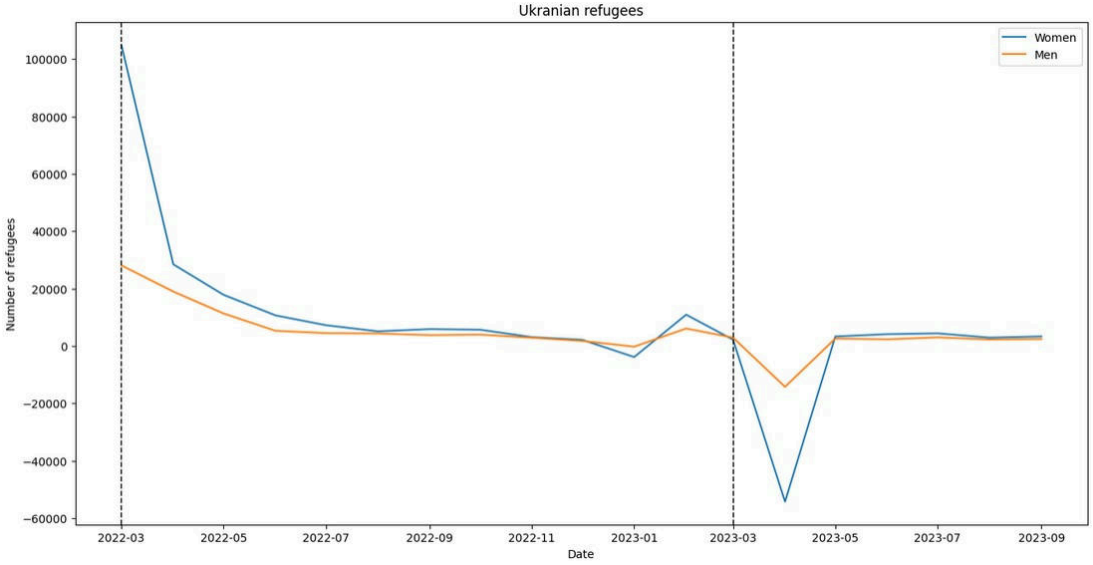

Figure 4: Time plot of refugees: *m_do_65* and z_do_65

5.2. Seasonal Plots

Finding recurring patterns is not a simple task, and relying on time plots alone can be misleading. It is better to pair it up with seasonal visualisation, which helps highlight recurring seasonal patterns. Our data had enough points and was clear enough for us to spot patterns just using the time plots, but we created this visualisation nonetheless to help support our claims.

Note that, once again, there are too many plots to display all of them here. We have provided some examples, while the rest is in our detailed analysis notebook inside our GitHub repository.

5.2.1. Jobseekers

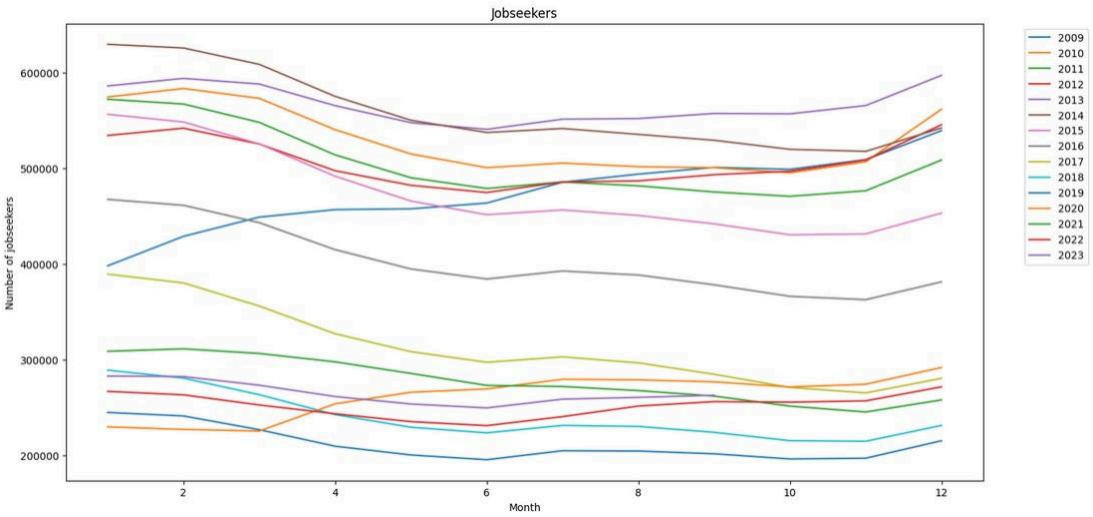

Figure 5: Seasonal plot of jobseekers: *uchazeciOZamestnaniUoZ*

Here, we can see much more clearly not only the downward trend from 2014 to 2020 but also the notably different course of 2009 and 2020 (and 2022 to a lesser extent). It is also very apparent that for the vast majority, the course is very cyclic with a strong pattern (= despite the lower values, the curve looks the same).

5.2.2. Already taken and closed positions

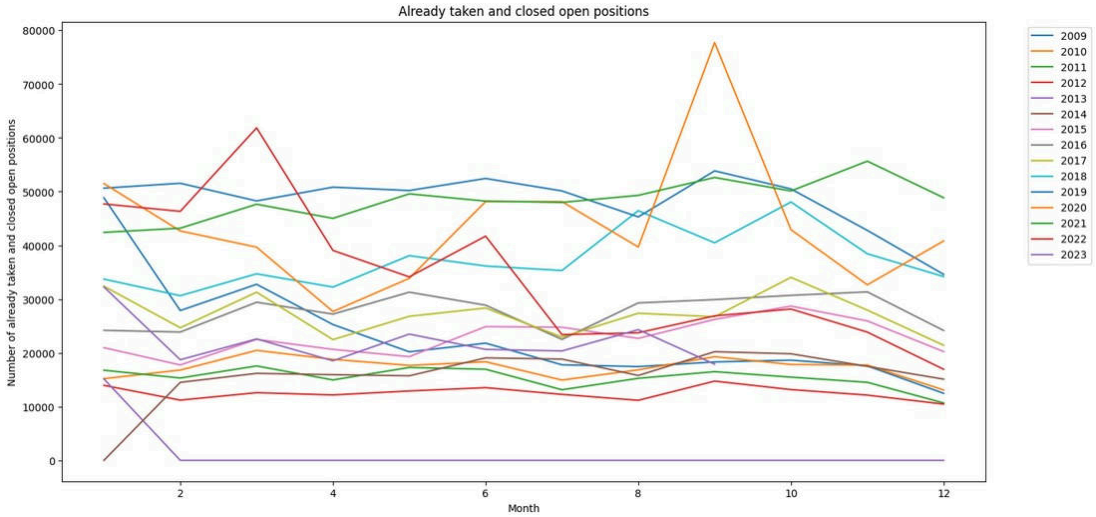

Figure 6: Seasonal plot of filled & closed positions: *obsazenaAZrusenaVPM*

We add this to draw contrast; there is no notable pattern or cyclic behaviour here, and we would be making a statistical error were we to assume otherwise. Negative answers in this test are just as useful as positive findings.

5.2.3. Seasonal findings

We came to the conclusion that some of the features are seasonal; thus, our model would benefit from using not only current values but also values of past years. We marked as seasonal the following predictors: *avg_monthly_salary*, *general_thefts*, *break_in_thefts*, *noveHlaseniUchazeci*, *absolventiSkolAMladistvi*, *noveHlasenaAUvolnenaVPM*, *obsazenaAZrusenaVPM*.

5.3. Correlation Plots

From a temporal perspective, our dataset is divided into two distinct categories: "long-term" data encompassing economic predictors such as minimum wage, average salary, gas prices, unemployment rates, etc., spanning the interval of the past 15 years (2009 – 2023). The "short-term" data is specifically related to the war, focusing on variables associated with the refugee situation in the Czech Republic. Consequently, a methodical examination of potential data correlations is undertaken in two sequential stages to enhance the precision of our analytical approach.

In the initial phase of our analysis, we calculated the correlation among all economic predictors utilizing the data spanning the preceding fifteen years. Three different methods were employed, some of which were robust with the aim of mitigating the impact of possible outliers through the years:

- The Pearson correlation
- The Spearman rank correlation (robust) – calculates the correlation based on the ranks of the data rather than the actual values
- The Kendall rank correlation (robust) – similar to Spearman's and particularly robust for small sample sizes

These calculations (no matter the method) confirmed a strong (positive) correlation in instances where a strong correlation was anticipated. Some specific examples would be:

- average salary and monthly minimum wage (τ = 0.87)
- monthly inflation rate and REER (τ = 0.82)
- general thefts and break-in thefts (τ = 0.92)
- number of unemployed people and number of newly unemployed people (τ = 0.86)

- newly opened/registered work positions and newly taken/closed work positions (τ = 0.93)

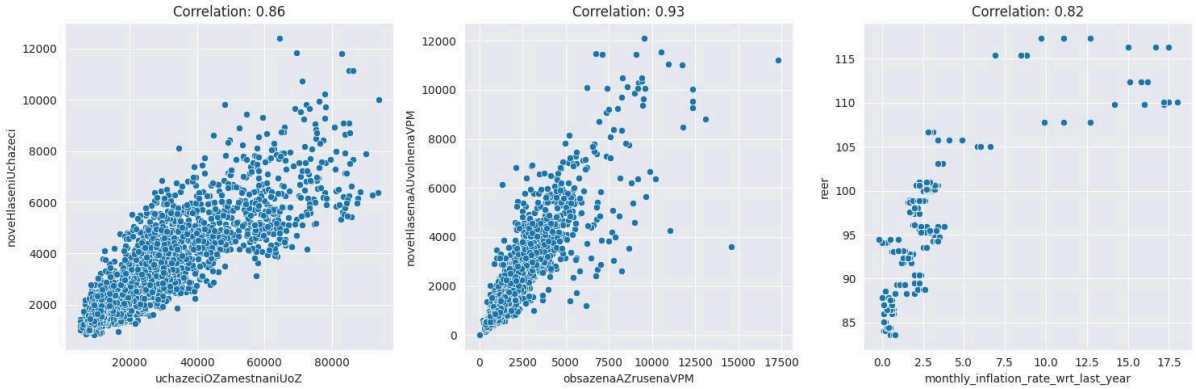

Figure 7: Examples of strong correlations between economic predictors, mainly ones related to unemployment. Correlation values as per the Kendall robust method

No strong negative correlation was unveiled. Moderate correlation, both positive and negative, was revealed in some cases, such as:

- monthly minimum wage and newly opened/registered work positions (positive, τ = 0.49)
- average monthly salary and newly taken/closed work positions (positive, τ = 0.54)
- number of newly unemployed people and general thefts (positive, τ = 0.55)
- newly opened/registered work positions and average gasoline price (negative, τ = −0.46)
- average monthly salary and youth/fresh graduates (negative, τ = −0.52)

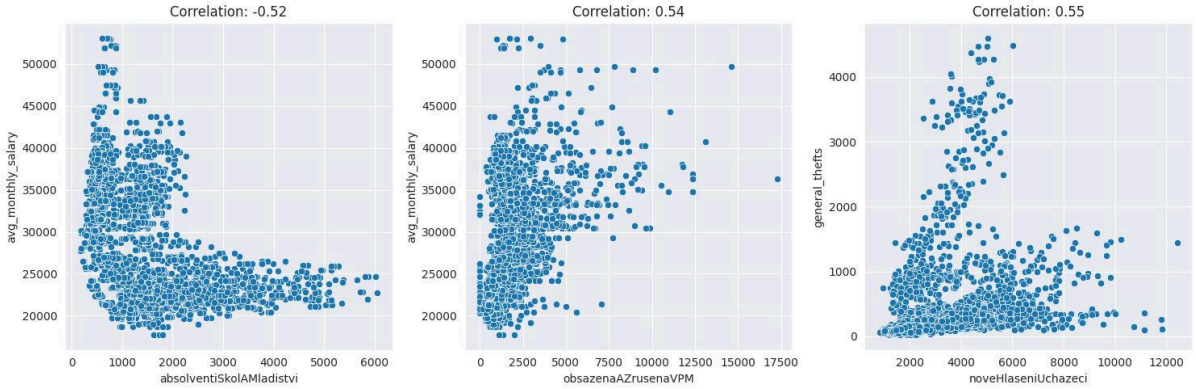

Figure 8: Examples of moderate correlation between economic predictors, mainly ones related to unemployment. Correlation values as per the Kendall robust method

In the subsequent phase, we focused on the temporal span encompassing the Ukrainian conflict (March 2022 – September 2023). For this interval, all predictors from the dataset were included, notably those pertaining to war refugees, which were entirely omitted in the previous stage. The identical triad of correlation methodologies was employed to compute correlation, and a discernible correlation between the overall number of refugees, irrespective of gender, and a myriad of predictors within the dataset was revealed:

- negative correlation:
  - REER, monthly minimum wage, import/export bilance
- positive correlation
  - newly opened/registered work positions, newly taken/closed work positions, general thefts, break-in thefts

It is imperative to note that the correlation observed between the number of refugees and the aforementioned predictors needs a cautious interpretation. The apparent associations may potentially be illusory and deceptive in nature. The observed fluctuations in the values of the specified predictors may seemingly coincide with the concurrent increase or decrease in the influx of refugees. However, such associations might be merely a consequence of the overarching impact of the ongoing war, exerting influence on all the enumerated predictors.

This potential confounding effect becomes particularly pronounced given the constraint of a notably brief time frame considered in this stage of analysis. The intricacies arising from the dynamic nature of conflict-related factors may contribute to a confounded interpretation of the observed correlations, emphasizing the need for circumspection in attributing causality or inferring substantive relationships in this context.

5.4. Lag Plots

Lag plots are another useful plot in time series analysis. They can show us whether the predictors have some autoregressive tendencies[10], help us detect outlier values or find seasonalities in the data. Lag plots, as described in [1], are plots where the horizontal axis is the original predictor value and the vertical axis is the value of the predictor shifted. This shift enables us to compare the value of the predictor with the value before some time (before one month, half year, year...).

For our case, we have decided to use not only lagged values but also cumulative lagged values. We use them for the refugee data to describe the total number of refugees that came during the last 2, 3, ... 19 months. To address these data, we introduce *cumulative lag plots* as a variant of the original lag plots where the difference is that on the vertical axis is now the cumulative value and not only the shifted predictor.

Overall, from the lag plots, we can conclude that, except few seasonal and autocorrelated predictors, the linear model should work well (and even better when we take the seasonality into account by including lagged predictors). On the other hand, the autoregressive models for time series would have probably worked poorly as the majority of the data seem to have random lag plots, and thus, the autoregressive model could not reliably predict the next value from the previous one.

5.4.1. Refugees Cumulative Lag Plots

In the upper left subplot of Figure 9, we can see that we have some possible outlier values in March and April, which corresponds to Section 5.1.4. Additionally, in the upper left subplot, we can see positive linear dependence and in the lower right subplot, negative linear dependence. These could cause autocorrelation issues when building a linear regression model. When we look at all subplots available at our GitHub we can also see that no seasonality occurred in these predictors (also corresponds to Section 5.1.4).

Almost the same also applies to *celkem* and *z_do_65* predictors whose lag plots are available on our GitHub repository.

5.4.2. Seasonal Patterns in Lag Plots

In Figure 10 we can observe that after 12 months, the plot shows less random and more linear dependency. This is evidence for the seasonality of this predictor (the same pattern also occurs after 24 or 36 months) and corresponds to findings in Section 5.1.3.

Another seasonality pattern was observed in predictor *absolvenntiSkolAMladistvi*. Here, the seasonality was weaker and also annual. Other possible seasonal predictors are *avg_monthly_salary* and *monthly_min_wage* where if the seasonality occurred, it would be monthly, which corresponds to Sec-

[10]For a definition, see [1]

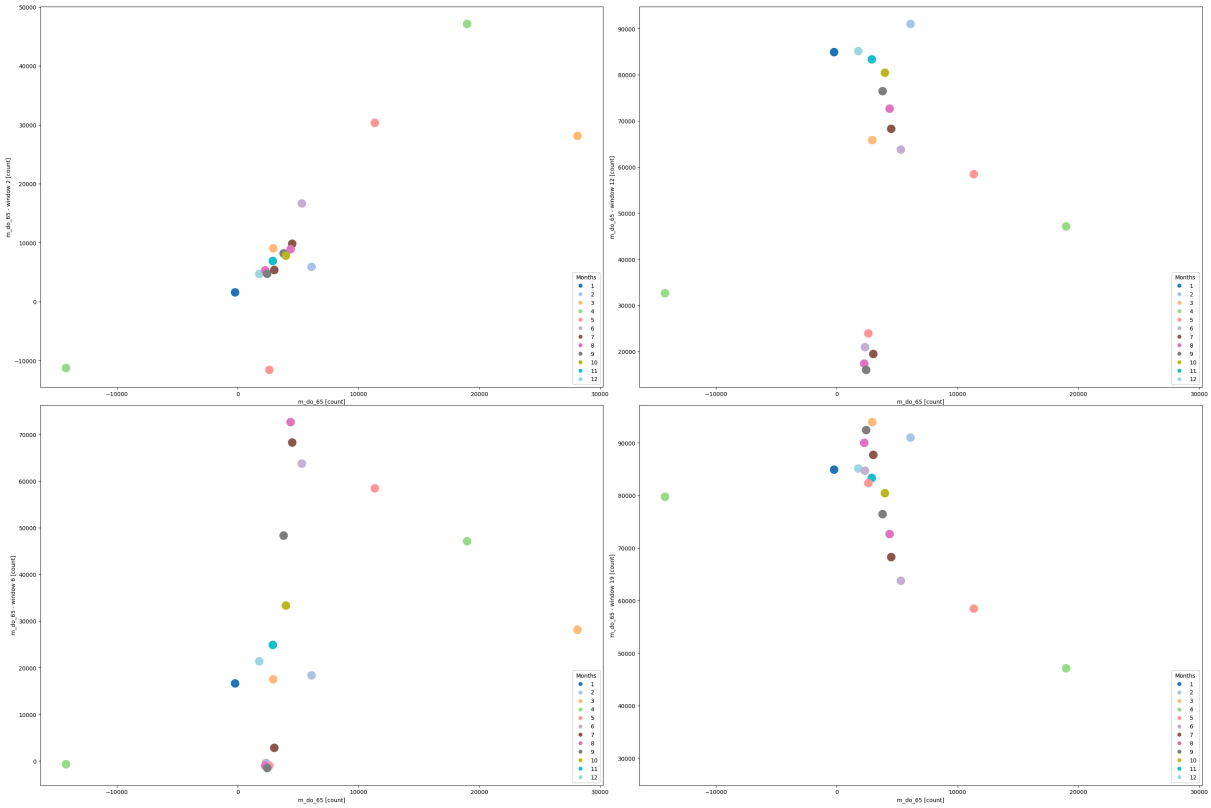

Figure 9: Cumulative lags of column m_do_65. Colours denotes the months

tion 5.1.1. In the rest of the predictors, no seasonality was observed. All evidence is available in our GitHub repository.

6. Models

The main objective of this project is to discern whether the recent influx of refugees has impacted the unemployment rates in the Czech Republic. Specifically, the project focuses on the female demographic due to the observed gender imbalance among migrants, as indicated by Figure 4. Those observations motivated the following question: "Do Ukrainian refugees significantly improve[11] the prediction of the Czech Republic's unemployment rate, or is that information redundant?". In order to address this question, two types of models were constructed and compared. Both of them are trained and predict outcomes based on a shared set of features comprising diverse economic indicators. However, their main difference lies in the incorporation of refugee-related information. The first model integrates data on the refugee situation, while the second model operates without consideration of such information. This comparative approach will allow us to determine whether the added information about the number of refugees results in a non-marginal improvement of the prediction. If the models behave comparably well, then refugees have no discernible influence over our targeted value.

Given the recency of the Russian-Ukrainian War, which commenced in 2022, the temporal span of features related to this conflict is notably truncated compared to the rest of the predictors. To address this temporal discrepancy, we considered two distinct methodologies:

- A: expand the interval of the shorter time series and set all pre-war refugee-related values to 0
- B: slice the longer time series and only work with the interval recorded for all features

We applied both solutions to both types of models, which resulted in four separate models:

[11] By improvement, we mean more precise predictions.

Figure 10: Lag plot of column *noveHlaseniUchazeci* (newly announced jobseekers)

- A1: Model containing both economic and refugee predictors, spanning over the last 15 years
- A2: Model containing economic predictors only, spanning over the last 15 years
- B1: Model containing both economic and refugee predictors, spanning over the last 2 years
- B2: Model containing economic predictors only, spanning over the last 2 years

Since we are working with time series data and aiming to predict a month's unemployment rate based on previous months, we are actually performing an *ex-post forecast*. As defined by R. J. Hyndman and G. Athanasopoulos [1], "*ex-post forecasts are those that are made using later information on the predictors. For example, ex-post forecasts of consumption may use the actual observations of the predictors, once these have been observed. These are not genuine forecasts, but are useful for studying the behaviour of forecasting models.The model from which ex-post forecasts are produced should not be estimated using data from the forecast period. That is, ex-post forecasts can assume knowledge of the predictor variables (the x variables), but should not assume knowledge of the data that are to be forecast (the y variable)*".

The preliminary analysis of the data aided in building models suitable for this particular task and dataset. Uncovering seasonal patterns meant using values lagged by one year, which made our model

somewhat autoregressive[12]. Because of the rather prominent presence of spikes in some of the models (e.g. Figure 4), we employed robust regression to combat possible outliers, and the revealed collinearities (Figure 7) implied that LASSO and Ridge regression would be beneficial. The scatterplots also exposed no apparent non-linear regressions.

Based on these findings, three regression techniques were implemented for the four models under consideration: Ridge regression, LASSO and Huber regression. Prior to inputting the data into the models, standardization was performed by subtracting the mean and dividing by the standard deviation. This normalization facilitates a more equitable comparison of the coefficients needed for regression with regularization penalty (e.g. Ridge, LASSO).

Thanks to our preliminary analysis we also knew some of the features exhibit strong seasonality, and so our models would benefit from having not only recent data but also past trends. For that reason, we constructed cumulative predictors for the features mentioned in Section 5.2.3 encompassing the whole course of the war.

7. Evaluation

Given the temporal nature of the dataset, the evaluation method had to be carefully selected to take into account this particular characteristic. As stated by R. J. Hyndman and G. Athanasopoulos [1], a suitable approach for training and evaluating models on temporal data is temporal cross-validation. In essence, the training fold encompasses values that precede those in the testing fold, essentially adhering to the principle of using past data to predict future observations (refer to Figure 11). Consequently, each step incrementally enlarges the size of the training set. Furthermore, we predict 14 values at once - one for each region of the Czech Republic - in order to avoid leakage of information (if we predict one region at a time, the fourteenth region will be using information about the rest of the regions).

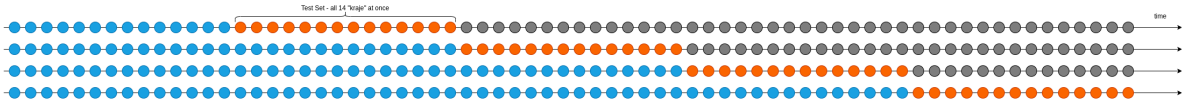

Figure 11: Depiction of temporal cross validation

Alternatively, we also experimented with a weighted testing technique, where newer information lowered the importance of older data (exponential decay), but that did not yield any improvement to our results, so we reverted to non-weighted temporal cross-validation in an effort to retain simplicity.

7.1. Results

After fitting each of the twelve models (three types of regression applied across four models as previously described) using the technique described above, we obtained very similar and comparable results for all three types of regression.

[12]Including the lagged features for the seasonal predictors truly boosted the performance of the models.

model	MAE	RMSE	R^2	$\overline{R^2}$
A1_ridge	1030.599	1153.345	0.956	0.955
A1_huber	860.548	993.639	0.967	0.966
A1_LASSO	1392.344	1639.836	0.911	0.910
A2_ridge	1411.005	1613.078	0.918	0.917
A2_huber	1202.265	1398.079	0.937	0.936
A2_LASSO	1757.902	1935.830	0.875	0.873
B1_ridge	436.909	526.488	**0.987**	0.984
B1_huber	485.229	577.741	0.986	0.984
B1_LASSO	**412.910**	**512.963**	**0.987**	**0.985**
B2_ridge	568.949	675.189	0.977	0.973
B2_huber	558.484	672.386	0.979	0.976
B2_LASSO	548.675	665.366	0.979	0.975

Table 1: Evaluation results of all types of regression on all four dataset versions. The best overall performance is highlighted in bold.

All of the metrics used for evaluation indicate that LASSO performed the best, and it was on the model containing both economic and refugee predictors spanning over the last two years (B1). Another advantage of LASSO is its ability to filter out irrelevant predictors, and three out of the four types of models contained such features (Table 2). Interestingly, for the short-term model, B1 *year* is apparently insignificant (not entirely unexpected), while the one-hot encoded feature *kraj* proved to be important for all models.

model	removed features
A1_LASSO	monthly_min_wage, monthly_inflation_rate_wrt_last_year
A2_LASSO	-
B1_LASSO	year, general_thefts, avg_monthly_salary, monthly_min_wage
B2_LASSO	general_thefts, break_in_thefts, avg_monthly_salary, obsazenaAZrusenaVPM

Table 2: An overview of the predictors LASSO filtered out during feature selection

The best overall model (B1) contained refugee information and outperformed both models trained on economic markers only (A2 and B2). The other refugee-based model, A1, is a long-term one and also outdid its long-term economic counterpart, A2. It is, however, understandable that it did not surpass the B2 model because of the severe imputation (inserting 0) on the dataset. Generally, the metrics suggest that adding the refugees features improved the quality of predictions of both the long-term and short-term models, which bodes well for our main question. The tests we conducted allow us to conclude that adding refugee-related predictors to the dataset does, in fact, lead to considerable improvements in predictions.

While all of the models mentioned above predicted the number of unemployed women, we also experimented with forecasting the *ratio*, rather than the absolute number, as it could potentially be even more telling of the prevailing trends. In order to do so, women's unemployment ratios were added to the dataset as features.
Upon fitting with Ridge regression and attempting to predict the next ratio, however, all coefficients were nearly zero. In our opinion, the fact that the previous ratios were around 50% sheds light on the possible cause. In that case, the model probably found it challenging to discern meaningful patterns

and simply predicted based on the learned previous ratios[13]. For that reason, we dismissed the ratio forecasts and stuck to absolute numbers.

7.2. Hypothesis testing

The task of comparing the two main types of modes and deciding whether refugee statistics boost the accuracy of unemployment rate predictions can be viewed as statistical hypothesis testing. In that case, our null hypothesis would be formulated as follows: "For both models, the metrics' values on each fold of the cross-validation have the same distribution[14]." Correspondingly, the alternative hypothesis states, "For both models, the metrics' values on each fold of the cross-validation do not have the same distribution.".

test	statistic value	p-value	df	
Paired t-test	−2.798	0.012	17	*
Wilcoxon signed-rank test	27.000	0.009	-	**

Table 3: Information about Paired t-test and Wilcoxon signed-rank test performed on RMSE values from cross-validation of models B1 and B2. One star stands for significance level 0.05, and two stars for 0.01

In order to be able to reject one of the hypotheses, two tests were performed: dependent t-test for paired samples and Wilcoxon signed-rank test[15]. The tests compared the best model with refugee information (B1) and the best model without (B2) based on RMSE[16]. Both tests[17] yielded p-values below the significance level (α) of 0.05 (see Table 3), which allows us to reject the null hypothesis and state that one of the models is legitimately better than the other one.

8. Conclusion

The project's primary goal was to investigate the relationship between the influx of Ukrainian refugees and women's unemployment rates in the Czech Republic. Various regression models, including Ridge, LASSO, and Huber, were employed, with meticulous consideration of the temporal nature of the dataset through temporal cross-validation.

A preliminary analysis of the data allowed us to gain a better understanding of the underlying trends, correlations and other specifics. Those findings facilitated the construction of models suitable for the specific dataset and, hopefully, getting more precise predictions.

The observed improvements in prediction accuracy when refugee-related features were included strongly suggest that the claim proposing no significant improvement in predictions with the addition of refugee-related predictors should be rejected. It was noted that incorporating refugees as features enhanced the predictive capabilities of the models, providing valuable insights into the dynamics between refugee influx and women's unemployment rates.

9. Contributions

We feel that every **Team A** member participated in equal measure towards our goal. We tried our best to achieve as much as possible and spent more than the allotted hours to complete the task thoroughly.

[13] This claim is supported by the intercept value.

[14] In other words, the metrics reported in Table 1 come from the same distribution for the compared pairs of models.

[15] It is a non-parametric version of the paired T-test.

[16] We also tested on MAE, R^2 and adj. R^2 - all of them gave similar results as RMSE with the exception of R^2 which had a p-value of 0.08 on the paired t-test.

[17] The t-test assumes a normal distribution, which is probably not true in our case, so we added its non-parametric version, the Wilcoxon test, which has no such assumptions and gave the same result (in the sense that we can reject the null hypothesis).

Time series went beyond the semester curriculum; thus, we spent time studying the principles behind it. Dataset construction had to be performed by us because we had to combine multiple sources, the workload of which we split. We went about preliminary analysis the same way. In multiple calls and meetings during the writing of this paper, we always discussed the next step in detail, ensuring everyone was on board. We dedicated a meeting to a discussion about the models themselves and their exact structure, after which two of us coded them (David Čech, Tomáš Mlynář), and two of us (*Korladinovi*) explained and examined them in this paper. Long live team A.

Bibliography

[1] R. J. Hyndman and G. Athanasopoulos, "Forecasting: principles and practice, 3rd edition". Accessed: Dec. 16, 2023. [Online]. Available: https://otexts.com/fpp3/