Final project report

Ostap Kmyta, Yurii Dmytruk

June 1, 2024

1 Introduction

A war is the most demanding challenge for any country's GDP [2], other economy-related aspects as well as for population's mental health [1] and other fields like education, safety and infrastructure. The Russian-Ukrainian war is of a particular interest since it is the most large-scale one in Europe since World War II.

In this paper we are observing the problem of resource allocation to revitalize Ukraine's economical and other indicators. We have decided to study the measures of landmine cleared area, total exports and percentage of educated population. Landmines affect country's agriculture since it is risky to cultivate mined land. Additionally, it slows down real estate renovation and causes post-war casualties. Export is a main source of foreign currency influx, which generates additional GDP, increases population's wealth and supports exchange rates of local currency. Finally, percentage of employment is a vital indicator as it shows country's ability to return profits, growing its economy. Thus, such objectives were chosen to investigate possible optimal solutions.

The main idea was to collect and analyze data on how the investments into the field affect it and solve the resource allocation problem in order to achieve balance among revitalization of these prosperity indicators.

2 Problem formulation

Our resource allocation problem [3] focuses on effective investing into revitalization of each issue. Given the sum N of money, the aim is to get an optimal solution to maximize the outcome of the combined investments. To solve the problem, we formulate the problem as three objective functions and one constraint, where

- $f(x_1)$ - total mined area cleared, sq km
- $f(x_2)$ - total exports, billions of US$
- $f(x_3)$ - employed population, %
- x_1 - money invested in mine clearance, US$
- x_2 - money invested in export revitalization, billions US$
- x_3 - money invested in education, billions US$
- $x_1 + x_2 + x_3 <= N$ - total money, billions US$

Since we don't have the explicit functions, we had to collect data in order to build surrogate models. This step influenced the initial choice of the objectives as we had not been able to find any data on the objectives we craved the most. The process of collection and processing of the data is described in detail in the next chapter.

3 Data collection and processing

Having started with a simple Google search, we came across multiple web pages, which offered different kinds of data for studying our case, however, none of four we looked at provided the data relevant enough to be selected for out work. After several tries of combining various datasets, we came across The World Bank website. This data bank provides numerous diverse datasets and quantitative reports on different topics. Among the list of topics we were able to find the datasets which were relevant to modelling our objective functions:

- Landmine Contamination Casualties and Clearance (LC3D) (used for *total mined area cleared* objective)
- World Development Indicators (used for *total exports* objective)
- Millennium Development Goals (used for *employed population* objective.

We started from analyzing how we can get our hands on the exact data we needed. As The World Bank website has functionality to choose series of data, we picked relevant data by choosing corresponding series only (Fig. 1).

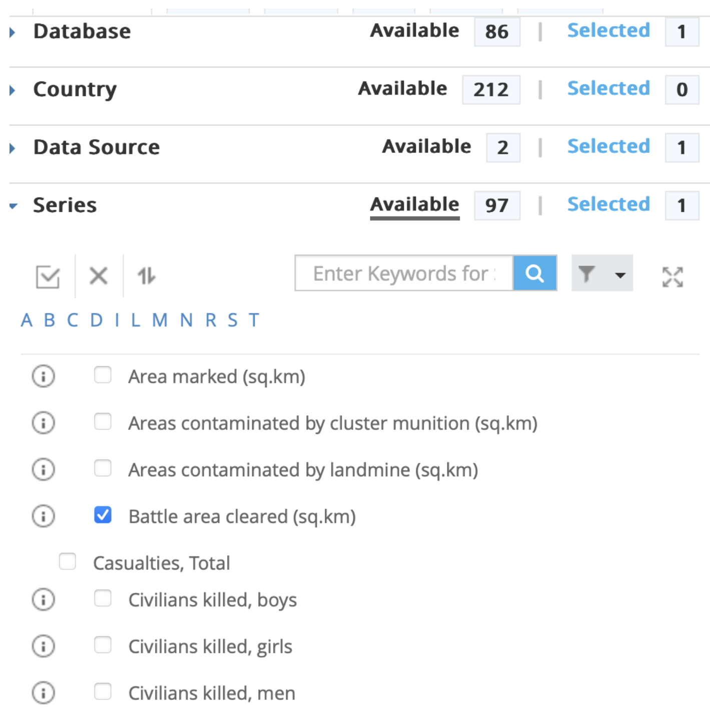


Figure 1: Filtering data on The World Bank Website

After a rough inspection, it was obvious that the *Landmine Contamination Casualties and Clearance* dataset is the smallest one, incorporating data from 1999 to 2012 only. This was the biggest limiting factor when intersecting data-series although we were satisfied with that. Right after choosing the data, we intersected the table with a total of 40 countries. To do so, we decided to filter countries we would be using information from, to the ones which are economically and resource-vise similar to Ukraine. This sorting process left us with 40 countries as data-points.

After downloading three datasets, we got three raw files (Fig 2) and we also found that the task was more difficult than we expected it to be as the datasets had multiple problems.

International Financial Assistance to	LCM.85	Latvia	LVA	0	6000000	12600000	4200000	5900000	9700000	6300000	68800000	28300000	0	0	0	0
International Financial Assistance to	LCM.85	Luxembour	LUX	0	0	530000	0	230000	80000	0	50000	50000	0	0	0	0
International Financial Assistance to	LCM.85	Marshall Isl	MHL	0	2160000	720000	600000	600000	390000	120000	190000	460000	0	0	0	0
International Financial Assistance to	LCM.85	Morocco	MAR	0	17000000	15100000	16900000	15250000	12000000	10000000	6200000	3500000	0	0	0	0
International Financial Assistance to	LCM.85	New Zealan	NZL	0	0	4720000	0	0	3980000	3500000	5700000	4500000	0	0	0	0
International Financial Assistance to	LCM.85	Peru	PER	0	0	0	0	0	0	0	0	0	0	960000	2534825	1996894
International Financial Assistance to	LCM.85	Romania	ROU	0	0	0	0	0	0	0	110000	0	0	0	0	0
International Financial Assistance to	LCM.85	Zimbabwe	ZWE	0	0	620000	0	0	0	0	0	0	0	0	0	0
International Financial Assistance to	LCM.85	Yemen, Rep	YEM	0	0	5900000	5600000	3630000	2640000	2460000	4070000	1100000	0	1000000	0	0
National Financial Assistance to Mine	LCM.83	Afghanistar	AFG	0	0	0	0	0	0	0	0	290000	0	0	0	3900000
National Financial Assistance to Mine	LCM.83	Albania	ALB	0	0	0	1580000	0	0	0	230000	240000	0	300000	100000	400000
National Financial Assistance to Mine	LCM.83	Angola	AGO	0	0	0	0	0	0	3000000	2500000	0	0	0	29183023	59600000

(a) Raw mine contamination data

Country Name	Series Name	1999	2000
Albania	Exports of goods and services (current US$)	540977914,6	700083398,1
Albania	Commercial service exports (current US$)	281600000	440300000
Albania	Merchandise exports (current US$)	351000000	261000000
Albania	Agricultural raw materials exports (% of merchandise exp	4,623562122	5,970691794
Albania	Foreign direct investment, net inflows (% of GDP)	1,282641334	4,108775955
Albania	GDP (current US$)	3212121651	3480355258
Armenia	Exports of goods and services (current US$)	383089746,9	423629630,5
Armenia	Commercial service exports (current US$)	132537724,6	133766597,2
Armenia	Merchandise exports (current US$)	232000000	294000000
Armenia	Agricultural raw materials exports (% of merchandise exp	3,732436419	4,898266077
Armenia	Foreign direct investment, net inflows (% of GDP)	6,612580809	5,450433156
Armenia	GDP (current US$)	1845482173	1911563669
Azerbaijan	Exports of goods and services (current US$)	1281380721	2118053577
Azerbaijan	Commercial service exports (current US$)	236496000	234444000
Azerbaijan	Merchandise exports (current US$)	929000000	1745000000

(b) Raw exports data

	1999	2000	2001	2002	2003	2004	2005	2006
Albania	48,289	48,45	48,035	48,413	48,215	47,994	47,791	47,742
Angola	59,252	59,235	59,264	59,231	59,339	59,303	61,038	62,847
Armenia	50,424	49,698	49,077	48,357	47,776	47,277	46,902	46,595
Azerbaijan	55,47	55,842	56,421	56,989	57,465	58,288	58,813	59,354
Bosnia and	36,228	36,559	36,247	35,478	35,104	34,384	34,178	33,557
Cambodia	76,952	76,66	81,746	80,996	80,113	78,685	79,888	81,096
Chile	50,275	49,486	49,483	49,252	49,738	50,905	50,903	51,712
Colombia	47,731	48,164	52,254	52,659	54,338	55,409	57,37	57,581
Croatia	46,637	45,03	44,859	44,421	45,045	46,183	46,582	46,851

(c) Raw % of employed population data

Figure 2: Raw dataset files

First of all, the website provides data in a very weird format (Fig 2a, 2b), which required a lot of manual processing. Secondly, the data in those datasets was not complete, and numerous cells were empty. Finally, when looking on the data, we were not sure, how to use the points to train the models.

3.1 Insufficient data

Initially, we have chosen a list of forty countries that boasted an acceptable amount of data points, however, when we started processing three Excel sheets, it was required to drop some of the countries, which lacked data in other datasets. We ended up dropping Afghanistan, Chad, Dem. Rep. Congo, Eritrea, Kosovo, Marshall Islands, Serbia and Somalia, shortening our list to 32 countries only. Additionally, we dropped column of year 2012 due to its emptiness.

Afterwards, when we understood based on which data columns (countries and years) we would be intersecting the data, a handful of empty cells were filled. Empty cells were filled with 0 in the mine clearance dataset, because it felt a right thing to do (no data - no land cleared). If to talk about investments-exports data, we had to ***come up with some data*** that would fit general trends.

3.2 Processing

We spent a lot of time refining the data. For datasets one and two, we built an intermediate one (Fig. 3).

Cleared (sq. km)

	1999	2000	2001	2002	2003	2004	2005	2006	2007
Albania	0	0	0,6	0,64	0,46	0,62	3,6	1,02	0,94
Angola	0	0	0	4,98	7,06	19	21,34	24,58	9,8
Armenia	0	0	0	0,02	0,3	1,08	0,22	0,44	1,04
Azerbaijan	0	0	9,54	1,4	2	9,54	12,38	9,7	15,08
Bosnia and H	0	13,1	14,22	11,08	12,66	13,34	8,4	8	6,6
Cambodia	31,12	25,06	24,36	30,8	35,2	34,6	37,8	61,6	70,8
Chile	0	0	0	0	0	0	0,2	0,04	1,46
Colombia	0	0	0	0	0	0	0	0,02	0,68
Croatia	28,34	28,66	19,62	25,38	61,7	63,66	67,36	19,8	49,54
Cyprus	0	0	0	0	0	0,08	0,46	2,96	1
Ecuador	0	0	0	0,02	0	0,04	0,02	0,02	0,02
Ethiopia	0	0	0	0,8	0	0	18	22	22,84

(a) Processed area cleared

Mine clearance Investments, $

	1999	2000	2001	2002	2003	2004	2005	2006	2007
Albania	0	0	2200000	4380000	3600000	3070000	5300000	2530000	1440000
Angola	17400000	0	13500000	21200000	21300000	28000000	38770000	50610000	19790000
Armenia	0	0	850000	4600000	250000	70000	0	40000	0
Azerbaijan	2200000	110000	5800000	4100000	5900000	0	0	0	1330000
Bosnia and H	4500000	11100000	16600000	20860000	17880000	2500000	16760000	31260000	30800000
Cambodia	20070000	9200000	21000000	27560000	17570000	41790000	25700000	30730000	31950000
Chile	0	0	0	0	0	2500000	2000000	3700000	1620000
Colombia	0	160000	50000	0	950000	2500000	2540000	5400000	10100000
Croatia	27000000	25380000	21700000	39240000	33600000	2500000	41910000	48800000	51500000
Cyprus	0	0	0	0	0	2500000	1000000	2260000	490000
Ecuador	3200000	2500000	3300000	1000000	0	2500000	770000	350000	610000
Ethiopia	330000	0	1900000	8370000	2500000	2360000	2600000	7900000	5800000

(b) Processed mine clearance investments

Figure 3: Intermediate dataset for mine clearance

3.2.1 Mine dataset

For mine clearance Excel sheet, we had *Mine area cleared*, *Total area cleared*, *Cluster munition area cleared* and *Battle area cleared* columns for clearance as well as *National* and *International financial assistance to Mine action* columns for investments. Using Excel, we summed these clearance and investment columns, obtaining only two columns, containing more generalised information.

3.2.2 Exports dataset

To obtain data table for investment - exports relation, we extracted 6 data columns:

1. Exports of goods and services (US$) ($E_{g\&s}$)
2. Commercial service exports (US$) ($E_{comm}$)
3. Merchandise exports (US$) E_{merch}
4. Agricultural raw materials exports (% of merchandise exports) (E_{agr})
5. Foreign direct investment, net inflows (% of GDP) (I)
6. GDP, (US$) ($GDP$)

Then we calculate total exports as:

$$E_{total} = E_{g\&s} + E_{comm} + E_{merch} * E_{agr}/100$$

and total investments as

$$I_{total} = GDP * I/100.$$

year	sq km of land cleared average	clearance investments average	foreign direct investment, US$ average	exports average, US$	Employment %, average
1999	2,87125	2631875	1433715017	16948208527	54,31840658
2000	3,0175	2884062,5	1159808014	19119611996	54,40715623
2001	3,50625	4501250	1175356954	17894958651	54,36053145
2002	9,4228125	6602187,5	1129326418	19129832524	54,65121889
2003	6,983125	6933750	1563016437	22303666942	54,8074373
2004	10,075	5309062,5	1974133396	28418759570	54,99665618
2005	9,75875	8464062,5	2368421939	33158263778	55,27790618
2006	16,456875	12219687,5	3392137400	39234475367	55,8148129
2007	14,53375	10795312,5	3825695914	47175449732	56,35165644
2008	0,04625	42187,5	4012999985	55206732055	56,47546864
2009	19,025	10530625	5184881663	46101744215	56,16431236
2010	21,52427669	9037420,125	4804351692	55831621869	56,22449982
2011	24,41410274	13432490,66	4534241863	67476528968	56,3677814

Figure 4: Final dataset

3.3 Merging

Since the last objective's data did not require any processing, we moved on to combining it into one spreadsheet to be able to read it with *pandas* library into the data frame.

We merged everything into one final dataset (Fig. 4), averaging the data by countries and obtaining 13 data points in total. The final dataset consists of 5 data type fields: year (indicates the year, where the data point comes from), square km of land cleared average (indicates average amount of area cleared of mine contamination during specific year for all the selected countries), clearance investments average (indicates average amount of money invested in mine clearance during specific year), foreign direct investment, US$ average(indicates average amount of money investment injections into the countries studied during the specific year), exports average, US$ (indicates amount of money in US$ which was made by exporting in a specific year), employment % average (indicates employment percentage in countries studied in a specific year). Now, we are ready to carry out surrogate modelling.

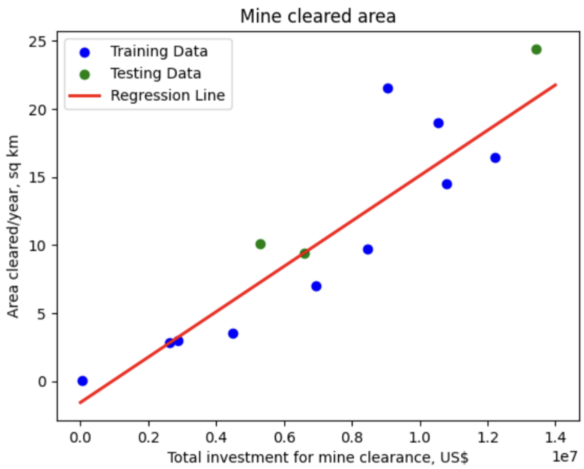
4 Surrogate modelling and optimization

Before I start describing the modelling and optimization process, I have to give a few remarks in the beginning.

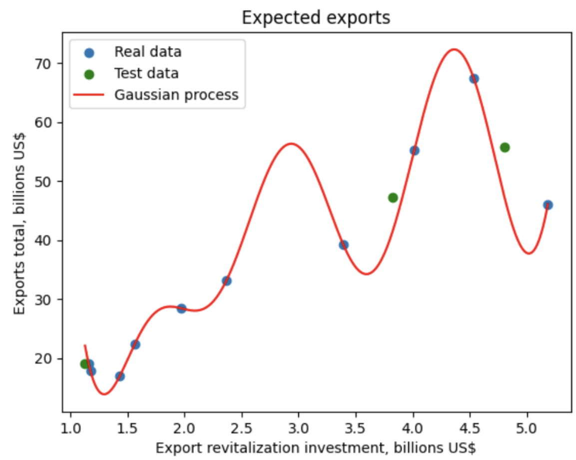
1. Investments range is the same in objective 2 and 3. This means we take *Foreign direct investment* (total) and assume that it is an investment into the specific industry, not a country as a whole. Additionally, we had scaled the metric to billions of US$.
2. Exports values were also scaled to billions of US$ before training to improve modelling.
3. The dataset of 13 entries was split into 80-20% train-test samples.

4.1 Modelling

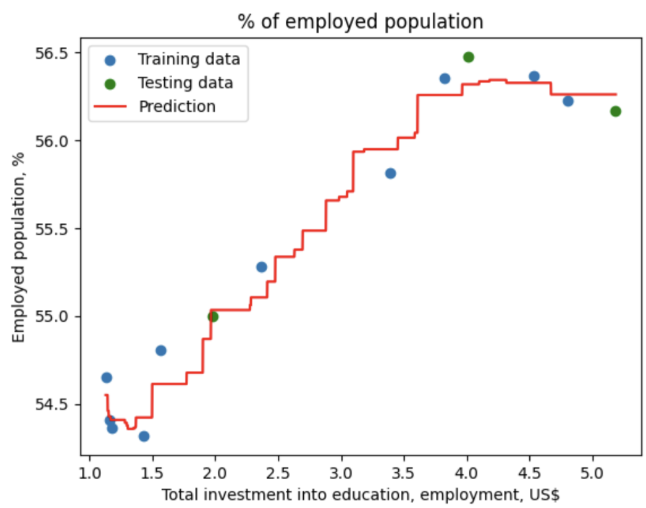
When playing with different models for our first objective, we looked on the plotted data and saw an approximate linear trend, so we made use of *scipy's Linear Regression*. This was the fastest and moderately effective solution, which showed a good result (Fig. 5a)



(a)



(b)



(c)

Figure 5: Modelling results

Modelling the expected exports was pretty difficult, because none of the modelling techniques yielded a good result at the beginning. Only after falling back to Gaussian Process and playing with some parameters were we able to obtain some closer results. However, for some reason the absolute best was achieved without specifying the covariance kernel function. We believe the Gaussian Process optimizer did some job of figuring the parameters during the fitting process (and it did better than we could), which lead us to this curve (Fig. 5b).

Finally, the % of employed population objective was modelled using *scipy's Random Forest Regression*. We tried using Randomized and GridSearch to figure out the parameters, but it did impair the R^2 and MSE scores, thus the default settings were chosen, leading us the following model (Fig. 5c).

Even though we have chosen these models, we are sure better results could have been achieved with other techniques and/or more rigorous parameter tuning. For example, as mentioned during the presentation, the K-fold cross-validation could have been used, which would have brought a great improvement (especially with a small dataset like ours). However, we did not manage to implement it in such a tight time frame, since the vast majority of the project time was spent on data preprocessing.

4.2 Optimization

To implement the optimization, we have chosen the Global Criterion method with the L_2 metric and z_{ideal} evaluated like this:

```
# Global criterion L2 scalarization
z_ideal = np.array([
    np.max([df['sq-km-of-land-cleared-average'].max(), \
    lreg.predict(model_one_range).max()]),
    np.max([df['exports-average,-billions-US$'].max(), gpr.predict(x_range).max()]),
    np.max([df['Employment-%,-average'].max(), tr.predict(x_range).max()])])

def global_criterion_l2(x, ideal):
  return np.linalg.norm(ideal - evaluate_models(x), ord=2)
```

The input value bounds were added to prevent models from extrapolating and the N constraint was handled in scipy format:

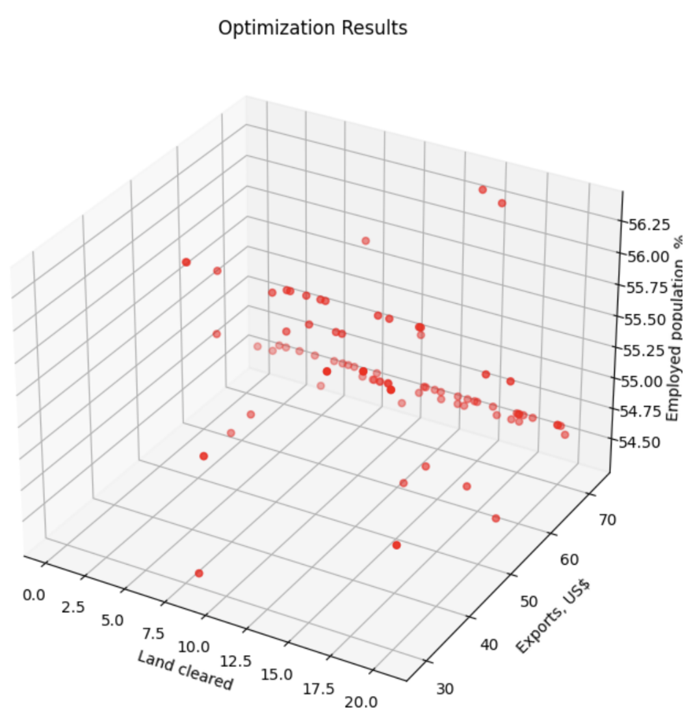
```
def total_money(x):
  return round(x[0] + x[1] * 1e9 + x[2] * 1e9) # obj 2 and 3 are in billions

N = 6e9 # max mony allowed - parameter
def constraint(x): # N - total >= 0
  return N - total_money(x)
```

With this code we ran optimization a hundred times choosing the initial optimizer point as a randomly chosen value within the x_i bounds. Here are some numeric and visual results:

```
x:[3.90481037e+06 4.36431209e+00 4.28376500e-01], Total money: 4796593403, y: [ 4.94338708 72.33119398 54.55003197], Scalarized value: 19.565686
x:[9.64464192e+06 5.34644249e+00 6.43459112e-01], Total money: 5999546242, y: [14.50362398 72.33369666 54.55003197], Scalarized value: 10.095786
x:[1.19580016e+07 9.06219345e-01 3.61555312e+00], Total money: 4533730470, y: [18.35674513 72.33092548 56.25675802], Scalarized value:  6.061305
x:[1.24905382e+07 5.34642650e+00 2.39369694e-01], Total money: 5598286733, y: [19.24373561 72.33019143 54.55003197], Scalarized value:  5.517246
x:[1.47282153e+06 2.93609003e+00 3.01128851e+00], Total money: 5948851362, y: [ 0.8926777  56.33589918 55.67736522], Scalarized value: 28.456023
x:[1.12999079e+07 4.38226330e+00 1.49861720e+00], Total money: 5892180413, y: [17.26062742 72.27248532 54.61292219], Scalarized value:  7.392208
x:[8.56825712e+06 1.87684555e+00 2.98359475e+00], Total money: 4869008555, y: [12.71080243 28.71748476 55.65687147], Scalarized value: 45.164109
x:[6.11791442e+06 4.36341897e+00 1.63045396e+00], Total money: 5999990853, y: [ 8.62952296 72.3312113  54.61292219], Scalarized value: 15.894088
x:[9.80013570e+06 5.34642937e+00 6.30252319e-01], Total money: 5986481821, y: [14.7626137  72.33081962 54.55003197], Scalarized value:  9.841674
x:[1.25517077e+07 2.93604032e+00 2.34478828e+00], Total money: 5293380307, y: [19.34561923 56.33589951 55.10548402], Scalarized value: 16.834999
x:[4.39204274e+06 4.35847675e+00 1.61081444e+00], Total money: 5973683230, y: [ 5.75491906 72.32633925 54.61292219], Scalarized value: 18.751913
x:[5.92314876e+06 4.36824583e+00 1.60789969e+00], Total money: 5982068676, y: [ 8.30512217 72.32784418 54.61292219], Scalarized value: 16.216299
x:[7.14250699e+06 2.93602254e+00 1.46825239e+00], Total money: 4411417434, y: [10.33607959 56.33589948 54.42220021], Scalarized value: 21.406949
x:[4.65753456e+06 5.34642379e+00 6.37535297e-01], Total money: 5988616617, y: [ 6.197121   72.3295961  54.55003197], Scalarized value: 18.318453
x:[1.17551679e+07 4.36793283e+00 1.62031201e+00], Total money: 6000000000, y: [18.01890624 72.32830617 54.61292219], Scalarized value:  6.660903
x:[2.94764177e+06 5.34643094e+00 1.08469423e-01], Total money: 5457848002, y: [ 3.34913147 72.33116408 54.55003197], Scalarized value: 21.152785
x:[9.75991730e+06 4.37308121e+00 1.61692814e+00], Total money: 5999769276, y: [14.69562613 72.31641188 54.61292219], Scalarized value:  9.895357
x:[5.16283042e+06 2.93604127e+00 2.24634205e+00], Total money: 5187546148, y: [ 7.03873946 56.33589951 55.03370244], Scalarized value: 23.660785
x:[7.51222410e+06 2.93605493e+00 3.05642176e+00], Total money: 5999988910, y: [10.9518787  56.33589947 55.7085721 ], Scalarized value: 20.920574
```

Figure 6: Some numeric results



It was satisfying to me how solutions with such different scalarized values are still considered to be Pareto optimal.

5 Conclusion

To sum up our work, I would like to start with some limitations. First of all, even though we had our best in finding and processing the data, in the end we understood that our approach is quite synthetic. First of all, we made up some data and made some assumptions regarding the investments. The real-world scenarios should be different. Secondly, we lack economic and possibly some geopolitical knowledge to work with such subtle . Numerous other factors may influence our objectives and these were not taken into account. Finally, we wish we had had more data to start with as it would have decreased uncertainty of our models, resulting in different Pareto front.

On the other hand, with this project we show that such kind of problems is relevant for modern days. It was a good exercise to practice on a real-word scenario, where the data had to be collected, processed and used by us as a future researchers and I believe we learned something for ourselves during the course.

References

[1] M. Carballo, A. Smajkic, D. Zeric, M. Dzidowska, J. Gebre-Medhin, and J. Van Halem. Mental health and coping in a war situation: The case of bosnia and herzegovina. *Journal of Biosocial Science*, 36(4):463–477, 2004.

[2] Jonathan Federle, André Meier, Gernot J. Müller, Willi Mutschler, and Moritz Schularick. Ökonomische folgen: Was kriege die welt kosten. *Wirtschaftsdienst*, 104(4):283–286, 2024.

[3] N. Katoh and T. Ibaraki. Resource allocation problems. In D. Z. Du and P. M. Pardalos, editors, *Handbook of Combinatorial Optimization*. Springer, Boston, MA, 1998.