RDM: Final Assignment

Gaja Nenadović

1 Description

The data we use in this assignment is the data from the Open Power Platform [1, 2]. It consists of multiple time series:

- actual and forecast values;
- for different countries;
- and for different variables (e.g. solar radiation, electricity price, temperature, load, etc.).

From their website:, "the time series data (`prices.csv`) contains data on Load, wind and solar, prices in hourly resolution. They are different kinds of timeseries data relevant for power system modelling, namely electricity prices, electricity consumption (load) as well as wind and solar power generation and capacities. The data is aggregated either by country, control area or bidding zone. Geographical coverage includes the EU and some neighbouring countries. All variables are provided in hourly resolution. Where original data is available in higher resolution (half-hourly or quarter-hourly), it is provided in separate files. This package version only contains data provided by TSOs and power exchanges via ENTSO-E Transparency, covering the period 2015-mid 2020. See previous versions for historical data from a broader range of sources" [1]. All data processing is conducted in Python/pandas and has been documented in the Jupyter notebooks linked below.

The weather data (`weather.csv` contains hourly geographically aggregated weather data for Europe. From their website: "contains radiation and temperature data, at hourly resolution, for Europe, aggregated by Renewables.ninja from the NASA MERRA-2 reanalysis. It covers the European countries using a population-weighted mean across all MERRA-2 grid cells within the given country" [2].

For simplicity sake, we will consider only the data for Austria, Bulgaria and Belgium for temperature (actual) and electricity load (actual and forecast) and only for the 60 minutes granularity and from a single provider (ENTSOE).

The variables we consider in the two data sets are:

1. AT_temperature: temperature weather variable for AT in degrees C;
2. BE_temperature: temperature weather variable for BE in degrees C;
3. BG_temperature: temperature weather variable for BG in degrees C;
4. utc_timestamp;
5. AT_load_actual_entsoe_transparency: Total load in Austria in MW as published on ENTSO-E Transparency Platform;
6. AT_load_forecast_entsoe_transparency: Day-ahead load forecast in Austria in MW as published on ENTSO-E Transparency Platform;
7. BE_load_actual_entsoe_transparency: Total load in Belgium in MW as published on ENTSO-E Transparency Platform;

8. BE_load_forecast_entsoe_transparency: Day-ahead load forecast in Belgium in MW as published on ENTSO-E Transparency Platform;
9. BG_load_actual_entsoe_transparency: Total load in Bulgaria in MW as published on ENTSO-E Transparency Platform;
10. BG_load_forecast_entsoe_transparency: Day-ahead load forecast in Bulgaria in MW as published on ENTSO-E Transparency Platform;

All timestamps are converted into standard format used in MySQL in `preprocessing.py` to weather_correct.csv and prices_correct csv.The two data sets, one with weather data and the other with time series about electricity related variables, are then transformed into long format with the code in the `preprocessing.py` to weather_long.csv and prices_long.csv to ease importation into database tables. They have the form of:

- utc_timestamp;
- fundament: variable considered, e.g. temperature or forecast for load;
- id;
- country;
- value: value of the said variable.

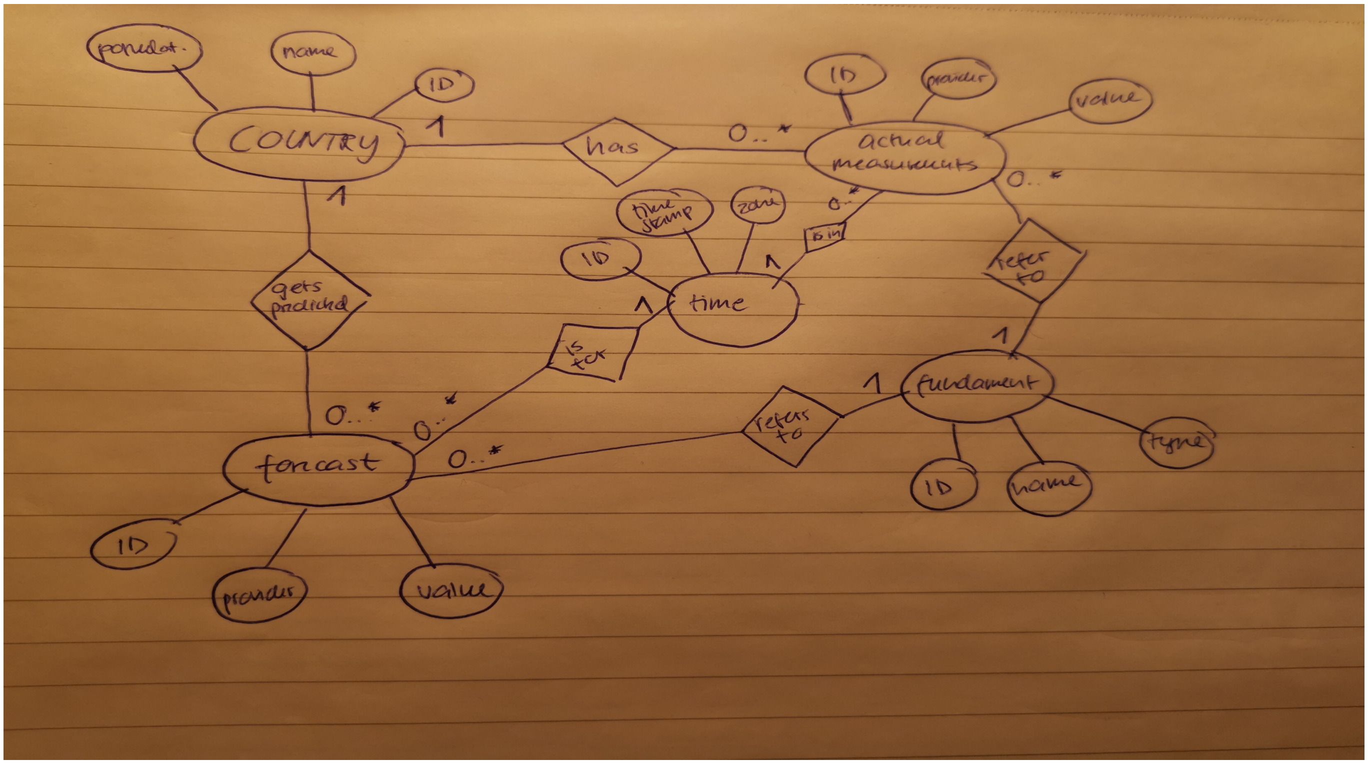
Research questions

The questions we will answer are:

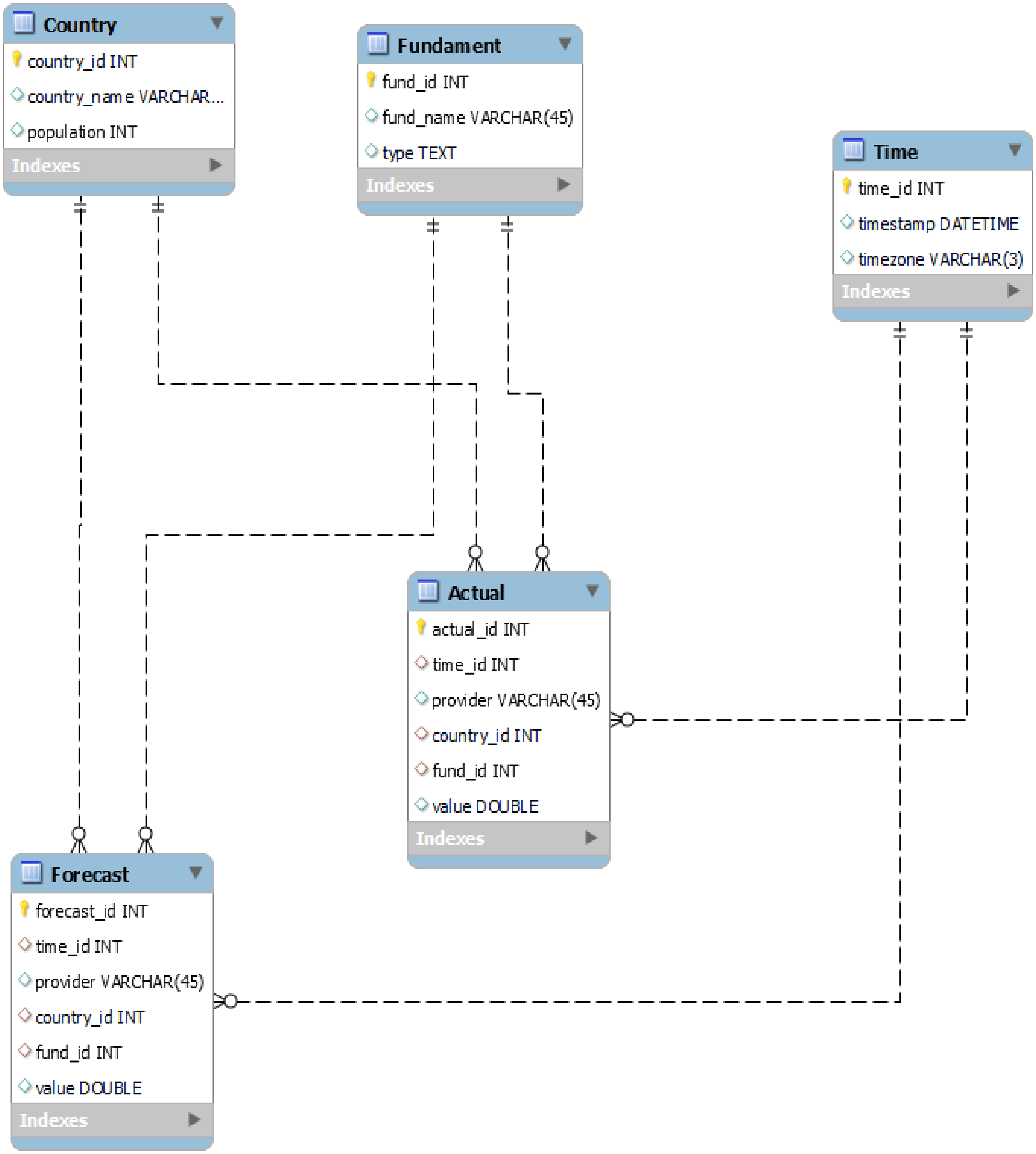
1. What is the average difference between forecast and actual load per country between January 1st and December 31st in 2016?
2. What is the trend in average yearly temperature between 1980 and 2020 per country?
3. what is the average temperature for countries starting with letter B between 1980-2020?
4. At which interval do the measurements of temperature become independent and are there differences between countries?
5. Which country has the highest average load over years 1980 to 2020 relative to is population?
6. Which year has the highest average load per country?

2 Models

Conceptual



Relational


Country
country_id INT
country_name VARCHAR...
population INT
Indexes
Fundament
fund_id INT
fund_name VARCHAR(45)
type TEXT
Indexes
Time
time_id INT
timestamp DATETIME
timezone VARCHAR(3)
Indexes
Actual
actual_id INT
time_id INT
provider VARCHAR(45)
country_id INT
fund_id INT
value DOUBLE
Indexes
Forecast
forecast_id INT
time_id INT
provider VARCHAR(45)
country_id INT
fund_id INT
value DOUBLE
Indexes

3 Analysis

All analyses and question to the answers are in the notebook file.

4 Reflection

Advantages of using a DB for such data is that is much more efficient in terms of time and accessing the data. If we had such data in separate csv file as the original data it would be very hard to access data for a certain country on a certain fundament, when you have data for the whole Europe on multiple variables in columns or even in separate files. It is also much slower to open and manipulate such huge sheets, while in mysql it was relatively fast to load the data and query it. It is also very easy to modify, e.g. remove, change or add values/columns, for example if we make a mistake or get additional data about an entitiy, e.g. a country in my case or a forecast.
The disadvantages are that this particular db, MySQL may not be appropriate for time series data and there exist special DBMS for such data. The thing is that we have multiple time series, which overlap in their timestamps but are linked to different entities and so the same timestamps must be repeated in the DB of my assignment (either repeated many times or each time series could have its own table - in any case not very efficient). If using specialized DBMS, this wouldn't be a problem. Another disadvantage is that the ER model might not fit the dynamics of the relationships for these data: it was hard to impose a relation between two entities sometimes.

References

[1] Open Power System Data. *Data Package Time series.* Primary data from various sources, for a complete list see URL, `https://data.open-power-system-data.org/time_series/2020-10-06`. 2020. DOI: `doi.org/10.25832/time_series/2020-10-06`.

[2] Open Power System Data. *Data Package Weather Data.* Primary data from various sources, for a complete list see URL, `https://data.open-power-system-data.org/weather_data/2020-09-16`. 2020. DOI: `doi.org/10.25832/weather_data/2020-09-16`.