First presentation of Producer prices workflow developed so far

20th October 2020

This is a first short presentation of the work performed so far. First of all, the document aims to present the actual structure of the theoretical workflow presented in previous presentations (figure 1). If the overall structure suits the technical team needs, methods used and other details can be discussed and customized since the tool is still flexible and simple enough to be easily changed.

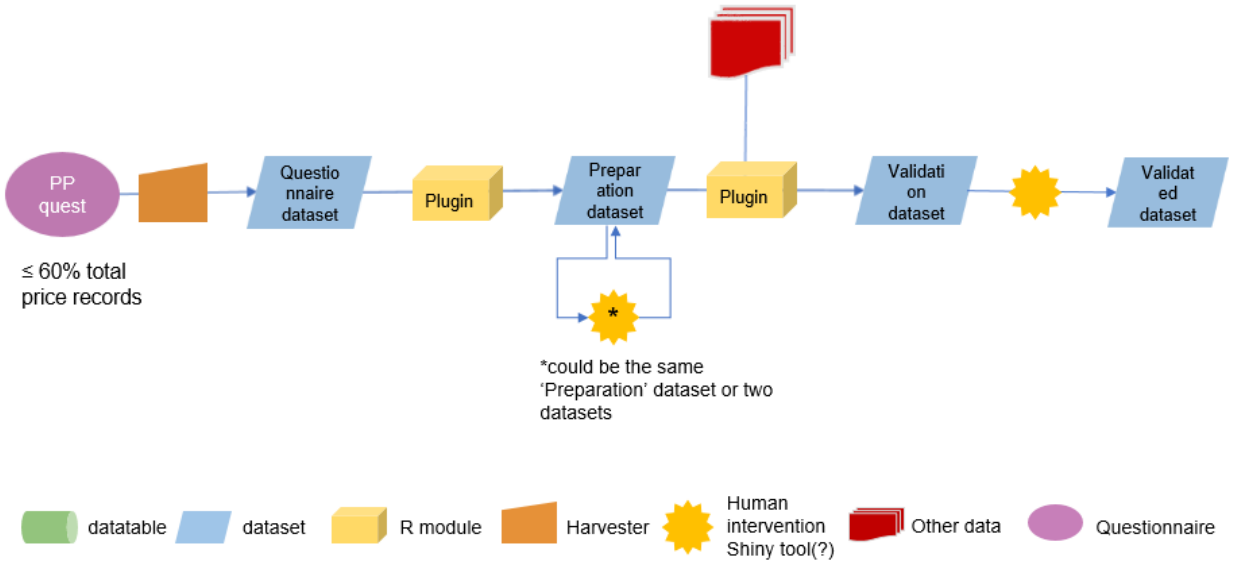


Figure 1 Overall workflow presented.

Plugin 1 workflow

The first plugin takes the data from the questionnaire dataset, i.e. in LCU, and fills a second dataset (*preparation* dataset) computing the price in USD and SLC and performing outlier detection.

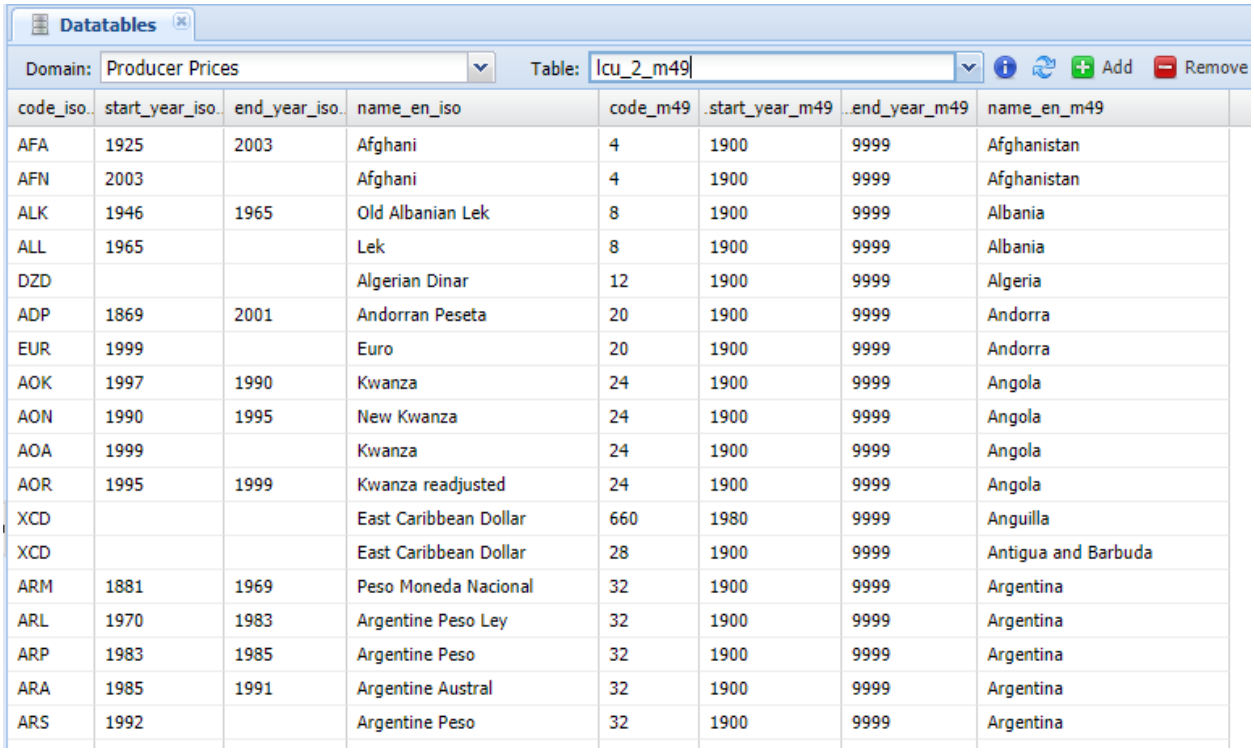
Since questionnaires data span over 3 years, the plugin checks if there are updates of previous years' figures. At the moment the plugin stores data differing for more than 0.1 from the existing value in a datatable that the user should revise. This is a new proposed step in the workflow and both threshold and revision method can be discussed.

Once the three units LCU, SLC and USD are available, the outlier detection process applies to the series. Currently a simple interquartile range is applied along with a time series approach (Friedman's super smoother). The outliers identified will be labelled with a specific flag and reviewed by the user in the

shiny app. If the detected outlier is confirmed by the user, the outlier value will be deleted and treated as a missing value, i.e. it will be imputed when the imputation plugin runs.

Plugin steps in detail:

- Parameter selection: countries and years (either start year chosen or by default last 10 years considered for outlier detection)
- Pull data from questionnaire dataset (last three years data in LCU)
- Pull country-currency datatable



Datatables

Domain: Producer Prices Table: lcu_2_m49 Add Remove

code_iso..	start_year_iso..	end_year_iso..	name_en_iso	code_m49	start_year_m49	end_year_m49	name_en_m49
AFA	1925	2003	Afghani	4	1900	9999	Afghanistan
AFN	2003		Afghani	4	1900	9999	Afghanistan
ALK	1946	1965	Old Albanian Lek	8	1900	9999	Albania
ALL	1965		Lek	8	1900	9999	Albania
DZD			Algerian Dinar	12	1900	9999	Algeria
ADP	1869	2001	Andorran Peseta	20	1900	9999	Andorra
EUR	1999		Euro	20	1900	9999	Andorra
AOK	1997	1990	Kwanza	24	1900	9999	Angola
AON	1990	1995	New Kwanza	24	1900	9999	Angola
AOA	1999		Kwanza	24	1900	9999	Angola
AOR	1995	1999	Kwanza readjusted	24	1900	9999	Angola
XCD			East Caribbean Dollar	660	1980	9999	Anguilla
XCD			East Caribbean Dollar	28	1900	9999	Antigua and Barbuda
ARM	1881	1969	Peso Moneda Nacional	32	1900	9999	Argentina
ARL	1970	1983	Argentine Peso Ley	32	1900	9999	Argentina
ARP	1983	1985	Argentine Peso	32	1900	9999	Argentina
ARA	1985	1991	Argentine Austral	32	1900	9999	Argentina
ARS	1992		Argentine Peso	32	1900	9999	Argentina

Figure 2: Country-currency datatable.

- Pull Exchange rates dataset
- Check that all currencies have been linked to their countries, if not error is returned.
- Link country to its currency
- Merge price with currency questionnaire dataset with Exchange rate dataset
- Convert LCU in USD
- Get previous series (preparation dataset or validated)
- If there is a break because of currency change, all the series is changed and converted into SLC, otherwise nothing changes
- Validated series and questionnaire data are merged
- The values overlapping (t-1 and t-2) are compared and if they differ for more than **e.g. 0.1** they are stored in a datatable (*revision2control*).
- Outlier detection on the SLC log prices according to two methods:

a. Interquartile range on the growth rates.
b. Automatic 'tsclean' function applied to logprices. 'tsclean' uses supsmu[1] for non-seasonal series and a periodic stl[2] decomposition with seasonal series to identify outliers.
c. Possible to add comparison with monthly data

- The identified outliers are marked as outliers on all elements (LCU, SLC and USD) and data are saved in the preparation dataset. Detected outliers will be manually revised with the shiny.
- An email is sent to the user

Notes: Change XR dataset

[1] Friedman's 'super smoother': supsmu is a running lines smoother which chooses between three spans for the lines. The running lines smoothers are symmetric, with k/2data points each side of the predicted point, and values of k as 0.5 * n, 0.2 * n and 0.05 * n, where n is the number of data points. If span is specified, a single smoother with span span * n is used. The best of the three smoothers is chosen by cross-validation for each prediction. The best spans are then smoothed by a running lines smoother and the final prediction chosen by linear interpolation. The FORTRAN code says: "For small samples (n < 40) or if there are substantial serial correlations between observations close in x-value, then a pre-specified fixed span smoother (span > 0) should be used. Reasonable span values are 0.2 to 0.4." Cases with non-finite values of x, y or wt are dropped, with a warning.

[2] Seasonal Decomposition of time series by LOESS

Plugin 2 workflow

The second plugin performs missing data and outlier imputation. It currently pulls data from the domains of Agriculture Production and Macro statistics and the Trade Openness Index (TOI) from a dedicated datatable.

After the last meeting, a set of methods to estimate missing value is being prepared. Currently, the ARIMA, the linear regression, Kalman smoothing, the price ratio (for specific commodities) and ensemble approach have been introduced. The covariate selection for the ARIMA and the linear model is performed via AIC.
The idea is either to manually review all the proposed method or to select the best model according to a common goodness of fit method.

Plugin steps in detail:

- Pull data from preparation dataset
- Expand to identify data to impute for the last year
- Pull data from Macro indicators (GDP and VA)
- Pull data from Agriculture production (Yield)
- Pull data from TOI datatable
- By country and commodity select the best model according to the AIC
- Forecast last year(s) missing value: ARIMA (computed within the plugin or pre-calculated model stored somewhere?) and linear regression, both with AIC based covariate selection, Kalman smoothing, price ratio and ensemble model.
  To be introduced: use of commodity group information, use of auxiliary information, linear interpolation, CPI & Food CPI, GDP and AgGDP deflators, Eurostat AgPPI Indeces, Exchange rates… The exact set of methods has still to be agreed on.

Shiny tool

The shiny tool has two main parts:

1. Outlier validation (figures 3 and 4)
2. Missing data imputation (figures 5 and 6)

Note: the example come from the data currently in the SWS dating from 2017. They have the suitable structure but no outlier or appropriate imputation example can be provided now.

Outlier validation tab

The outlier validation tab is rather simple. The user select from the Outlier list one product (e.g. Copra price in 2012 in Cambodia). The list is created starting from the outlier flagged by the first plugin. Currently, each year is considered as a separated object in the list (e.g. Copra in Cambodia has 3 points in the list for 2010, 2011 and 2012). This is the easiest solution but this can be changed and the same series can be shown in the same tab.

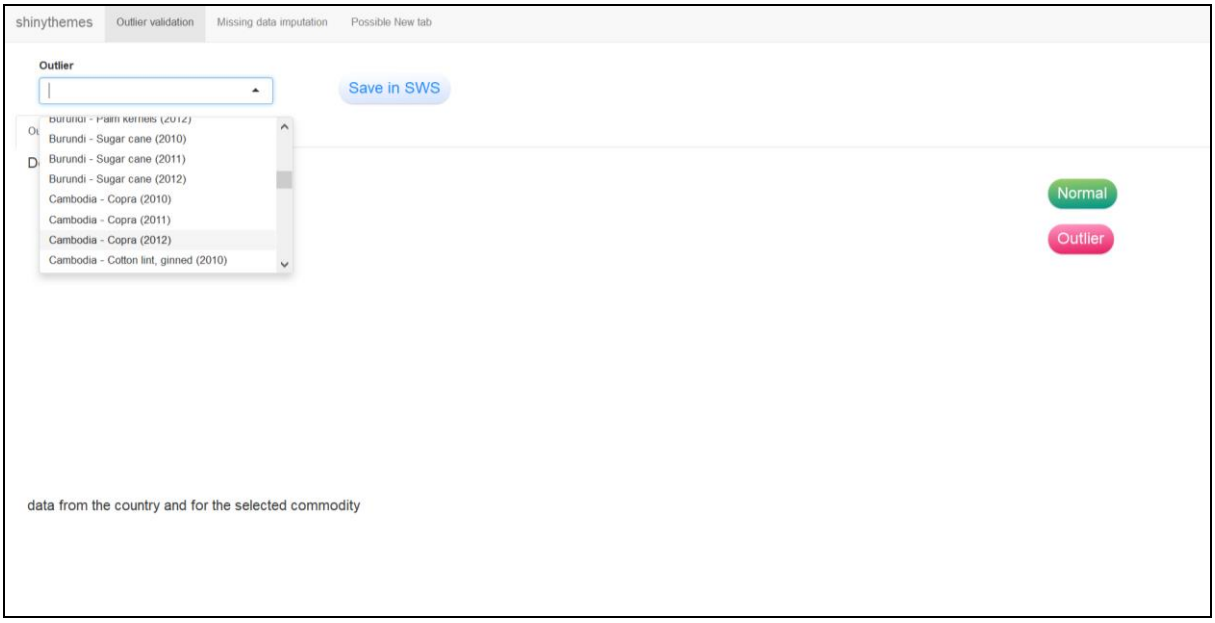


Figure 3 Outlier validation. Product selection.

Once the product is chosen the app shows the plot of the series with the outlier detected marked in red. Below, a table shows all the data for the selected country and all the available price for the chosen product. The table also reports the yearly price variation for each commodity so to help the user to check if the outlier detected depends on a shock in the country or on a specific product shock.

After these check the user can decide if the figure is actually an outlier or not by clicking respectively the button ‘Outlier’ or ‘Normal’. In case the ’Normal’ button is chosen the value is kept and flagged again as official figure and not as an outlier. If the figure is acknowledged as an outlier the figure is set as NA and will be imputed during the imputation phase.

Once one of the two buttons has been clicked, the commodity disappears from the list and the user can switch to the next one, until they are all checked.
In order to save the work the user can click on the top ‘Save in SWS’ button and transfer the changes to a SWS session.

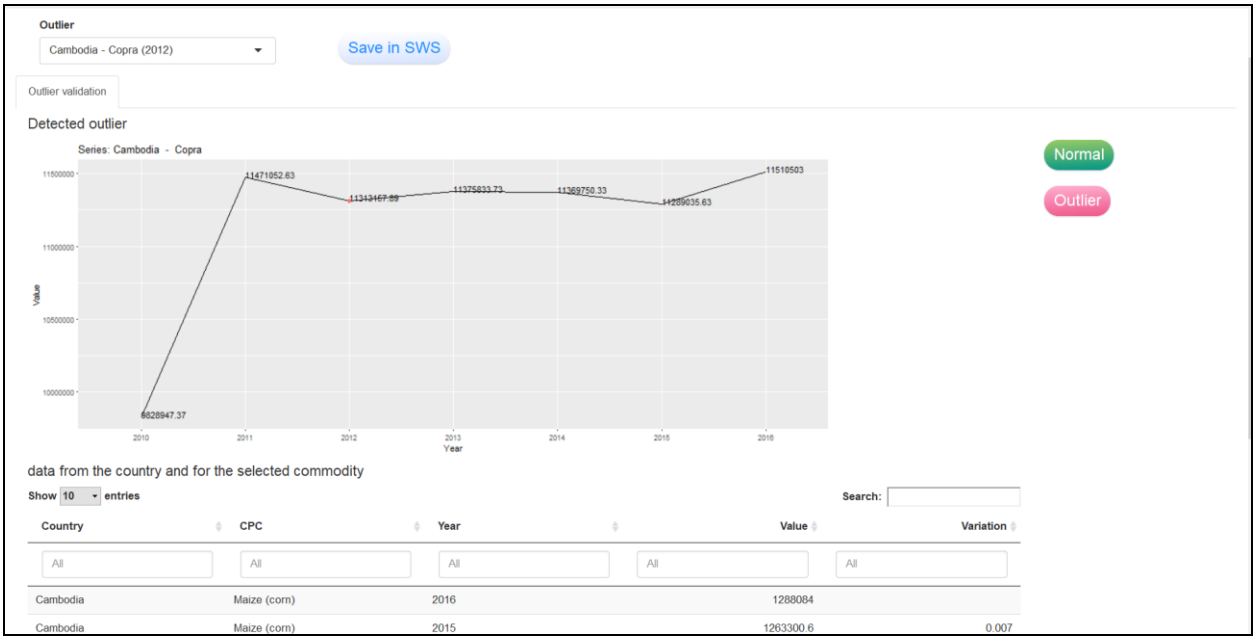


Figure 4: Outlier validation. Manual decision.

Missing data imputation tab

The imputation tab has a similar structure to the outlier one. The user chooses the product from the list and the shiny shows the product series and the tab with the country data and the product data. If the user accepts the proposed value(s) he can validate otherwise clicking on the refuse button he can insert manually a value (figure 6).

Figure 5: Missing data imputation. Imputation proposal.

By clicking the 'Refuse' button two other buttons appear. In one the user can insert the value to be imputed. With the 'Impute' button the user saves the manually imputed value into the dataset.

As in the previous tab, all changes have to be saved into the SWS though the 'Save in SWS' button.

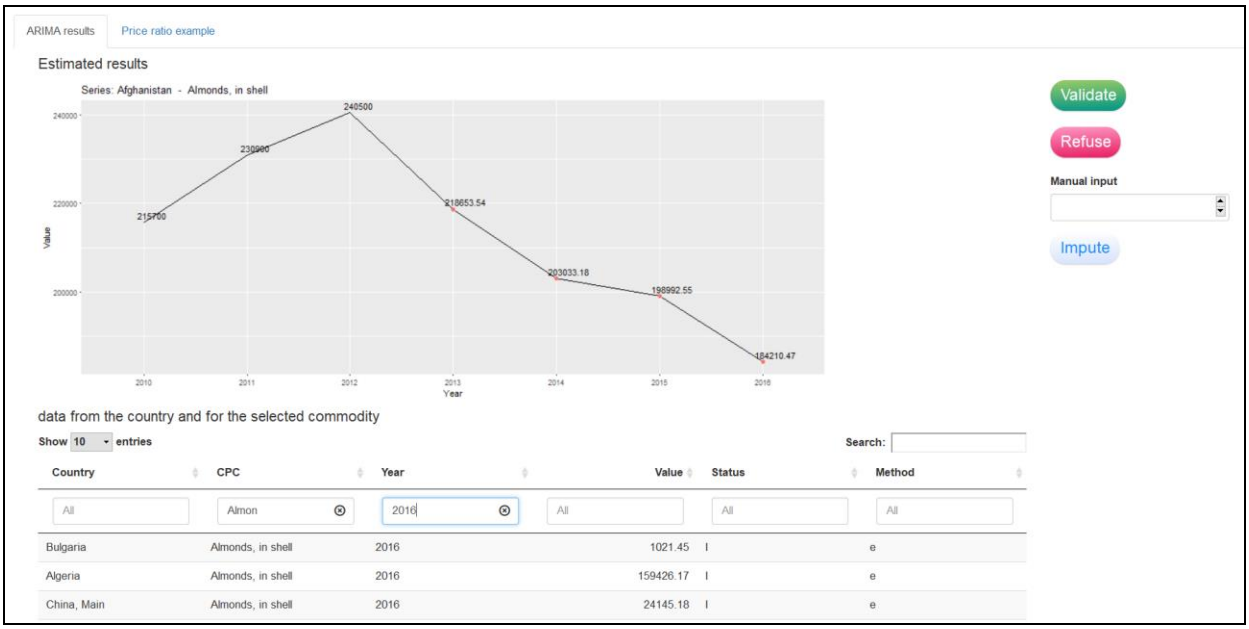


Figure 6: Missing data imputation. Manual imputation.

If more than one model has to be proposed to the user the validate button can be changed in a 'multiple choice' button with options shown in the plot. An example of it is proposed in figure 7. The figure refers to a tool developed for the Fishery unit and validated few month ago.

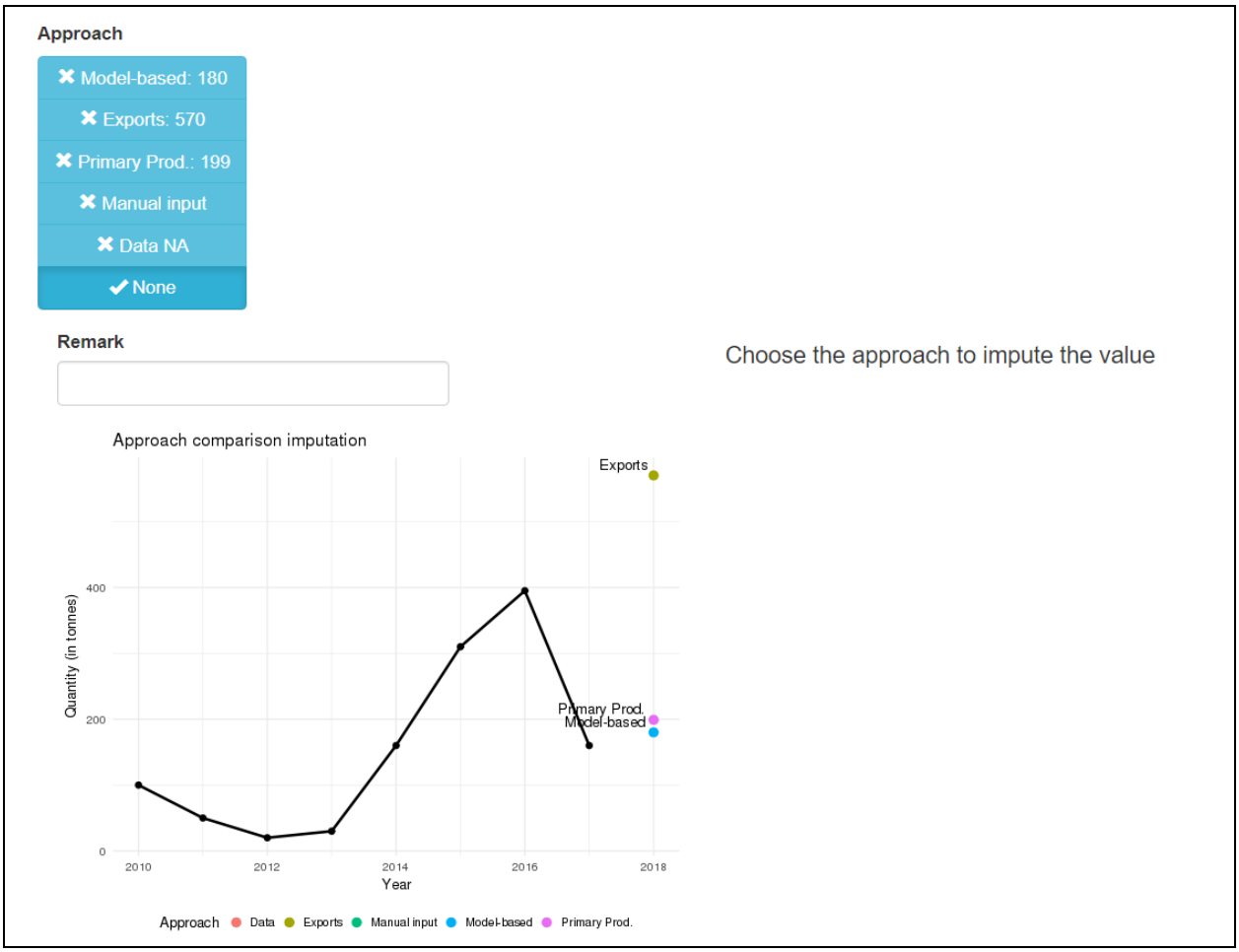


Figure 7. Example of different approach proposal to the user.