

JRC TECHNICAL REPORTS

INFORM: scientific and technical improvements in 2017

*Missing values
imputation and IT
developments*

Marín-Ferrer, M.

Doherty, B.

Béjar García, J.

Luoni, S.

Vernaccini, L.

2017



**INDEX FOR RISK
MANAGEMENT**

INFORM
INDEX FOR RISK MANAGEMENT

This publication is a Technical report by the Joint Research Centre (JRC), the European Commission's science and knowledge service. It aims to provide evidence-based scientific support to the European policymaking process. The scientific output expressed does not imply a policy position of the European Commission. Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use that might be made of this publication.

Contact information

Name: Montserrat Marin-Ferrer

Address: Via Fermi, 2479 – Ispra (VA), Italy

E-mail: montserrat.marin-ferrer@ec.europa.eu

Tel.: +39 0332 785810

JRC Science Hub

<https://ec.europa.eu/jrc>

JRC110540

EUR 29081 EN

PDF ISBN 978-92-79-77804-9 ISSN 1831-9424 doi:10.2760/076136

Ispra: European Commission, 2018

© European Union, 2018

Reuse is authorised provided the source is acknowledged. The reuse policy of European Commission documents is regulated by Decision 2011/833/EU (OJ L 330, 14.12.2011, p. 39).

For any use or reproduction of photos or other material that is not under the EU copyright, permission must be sought directly from the copyright holders.

How to cite this report: Marin Ferrer, M., Doherty, B., Bejar Garcia, J., Luoni, S. and Vernaccini, L., INFORM scientific and technical improvements in 2017: Missing values imputation and IT developments, EUR 29081 EN, Publications Office of the European Union, Luxembourg, 2018, ISBN 978-92-79-77804-9, doi:10.2760/076136, JRC110540.

All images © European Union 2018

Contents

- Abstract2
- 1 Introduction.....3
- 2 Current approach used in INFORM for addressing missing values.....4
- 3 Random Forest Regression applied to INFORM.....6
- 4 Results.....7
 - 4.1 Prediction performance.....7
 - 4.1.1 Score7
 - 4.1.2 MSE9
 - 4.2 Comparison with the current prediction method..... 12
 - 4.3 Areas for Further Research 16
- 5 Data management..... 17
 - 5.1 Inform tool: external access..... 17
 - 5.1.1 Integration with DNN 17
 - 5.2 INFORM web API 20
- 6 Conclusion..... 22
- Annex 1: Predicting missing Data in INFORM 23
- Annex 2: INFORM development report 38

Abstract

The European Commission Joint Research Centre (JRC) is the technical and scientific leader of the Index for Risk Management (INFORM), being the responsible for the development of its methodological improvements and their corresponding implementation.

This publication describes the major methodological and technical improvements on the INFORM model implemented by the JRC in 2017.

Although the indicators have been selected on the basis of their reliability, consistency, continuity and completeness, most of them do not cover all the countries with data for every year. This results in a significant number of missing values, irregularly distributed among countries, time and indicators.

This report describes an innovative approach for predicting missing values using the most advanced statistical technics, the so called machine learning, that has been combined with the traditional composite indicator adopted by INFORM in order to improve the accuracy of the risk index.

We also present the IT latest developments that support the INFORM model, including the web platform for managing the INFORM Subnational models and improvements in the new Application Programming Interface (API).

1 Introduction

The Index for Risk Management (INFORM) is a composite indicator that identifies countries at risk of a humanitarian crisis or disaster that could overwhelm their national response capacity. The INFORM index supports a proactive crisis and disaster management framework. The INFORM initiative began in 2012 as a convergence of interests of UN agencies, donors, NGOs and research institutions to establish a common evidence base for global humanitarian risk analysis.

The INFORM model is based on risk concepts published in scientific literature and envisages three dimensions of risk: Hazards & exposure, Vulnerability, and Lack of coping capacity. The INFORM model is split into different levels to provide a quick overview of the underlying factors leading to humanitarian risk and builds up the picture of risk using more than 50 core indicators.

The European Commission Joint Research Centre (JRC) is the technical and scientific leader of the model, being the responsible for the development of its methodological improvements and their corresponding implementation. INFORM partners organize an annual meeting where needs and gaps are shared and the strategic developments are discussed. The scope of this publication is to describe the INFORM methodological and technical improvements implemented by JRC in the 2017 following the discussions held with partners.

Although the indicators have been selected on the basis of their reliability, consistency, continuity and completeness, most of them do not cover all the countries with data for every year. This results in a significant number of missing values, irregularly distributed among countries, time and indicators.

In this report, we introduce an innovative approach for predicting missing values using advanced statistical technics, which will allow improving the accuracy of the index.

We also present the IT latest developments in support to the INFORM model, including the web platform for managing the INFORM Subnational models and the improvements in the new Application Programming Interface (API).

2 Current approach used in INFORM for addressing missing values

In the current version of INFORM, if data for some countries are not available for a given year, a systematic imputation of missing values is made using the data from the most recent year available over 5-years span. Only for two indicators in the Food Security component, namely 'Prevalence of undernourishment' and 'Average dietary energy supply adequacy', we use the regional average for imputing missing values.

In the case of missing data due to weak coverage, the approach is to introduce more than one indicator for the same component so that the indicators complement each other, taking the average index of the remaining indicators. This method is an implicit treatment of missing values, where for each unit only observed values are considered.

These are currently the only criteria used for imputing missing values in INFORM.

There are many aspects where missing values could influence the INFORM results:

- **Missing data can distort the real value of the composite indicator.** Missing data cannot be completely avoided. The goal of the composite indicator is to aggregate the different aspects of humanitarian risk. Whenever certain values are missing, the aggregation process fails as a tool to compensate a deficit in one dimension /category/components by creating a surplus in another. In the case of poor coverage, we introduce, whenever available, more than one proxy measure for the same component so that they complement each other.

Table 1. Countries with more than 20% of missing values in INFORM 2018 version

Country	Missing values (% of total)
Liechtenstein	22 (43 %)
Tuvalu	15 (29 %)
Nauru	14 (27 %)
Marshall Islands	13 (25 %)
Dominica	13 (25 %)
Saint Kitts and Nevis	13 (25 %)
Grenada	12 (24 %)
Democratic People's Republic of Korea	12 (24 %)
Antigua and Barbuda	11 (22 %)
Palau	11 (22 %)
Eritrea	10 (20 %)
Kiribati	10 (20 %)
Micronesia	10 (20 %)
Somalia	10 (20 %)
Libya	10 (20 %)

In INFORM 2018 39 countries have all data, while 15 countries have more than 20% of missing values (Table 1).

- **Countries in conflict.** In countries facing internal conflicts (e. g. Syria, Iraq and Libya), the reliability of the data (when available) is normally weak, or the data are out of date. Therefore the resulting INFORM score for those countries is not considered fully reliable.
- **Lack of real-time data.** Some indicators in the INFORM index are designed to reflect the real-time situation but there are time constraints that should be kept in mind. Firstly, there is a time lag between a situation changing and the indicator reflecting this change and, secondly, the indicators are usually issued with delays because they need to go through a validation process.
- **Trend analysis.** The historical results are back-calculated using the same methodology and data source of the published release. Incomplete historical values can strongly influence trend analysis.

Recent UN report¹ suggests using different methods, including Artificial Intelligence, to fill data gap.

In order to reduce the negative effects of missing values in the INFORM results, we present an advanced statistical methods to predict them (Chapter 3).

¹ Innovative Big Data approaches for capturing and analysing data to monitor and achieve the SDGs (2017), ESCAP.

3 Random Forest Regression applied to INFORM

INFORM, as along with many others, uses tools that extract current information, sift through data looking for patterns that are relevant to our problem and returns answers and error levels. The process of developing these kinds of tools has evolved throughout a number of fields including chemistry, computer science, physics, and statistics and has been called machine learning, artificial intelligence, pattern recognition, data mining, predictive analytics, and knowledge discovery (Trevor Hastie, 2009). While each field approaches the problem using different perspectives and tool sets, the ultimate objective is the same: to make an accurate prediction.

The main motivations for using these new methods in INFORM arise from the need to predict certain trends in countries for which this would otherwise not be possible due to the lack of information or parameters in the original data. To achieve this, the objective is to find the maximum correlation between available information and the indicators not present.

We have applied the most advanced statistical methods to the current INFORM development. We try to show how different mathematical methods can deal with missing data (Longford, 2005), going further than applying just imputation methods from existing indexes (predictors or variables). We will be able to predict values and fill the gaps in those indexes using the information available for each year and each country. These predictions come with a score or error level to provide a level of accuracy to the model.

The data used in INFORM come from different authoritative sources providing a great deal of knowledge for the report. However, due to its nature, the raw data comes with gaps, creating some noise in the final results. Current INFORM data sets include information from 20 (years) x 191 (countries) x 54 (indicators). Most of the information used comes from the last 5 years.

We have added new data sources from World Bank Development Indicators² and the World Health Organisation Global Health Observatory³ to this new approach. These databases provide a wide range of predictors which let us create several different data sets and therefore different models. The main reason for using these sources is to maximise the available information relative to each country and the correlation with the values we intend to predict. As a result of this union, the new dataset used has 67 (years) x 249 (countries) x 507 (indicators).

By adding more data to the data set we are adding more information to the predicted model and improving the future score of each indicator. The main goal of adding these data is not including them in the INFORM methodology, but rather increasing the probability of having a better score in an indicator that is in fact used in the methodology. Adding more data sometimes means adding more noise, and in this case the use of the Random Forest Regressor RFR works as a filter removing that noise to a large degree.

This new approach is based in the field of Supervised Learning an area inside Artificial Intelligence often called Non-linear Regression (Trevor Hastie, 2009). Among every available model for regression, we have tested different methods like Ridge, Lasso, ElasticNet or Random Forest, finding Random Forest the best balance between performance and complexity.

A Random Forest (Segal, 2004) is a meta-estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy while controlling overfitting. Overfitting is one critical problem that may make the results worse, but for a Random Forest algorithm, if there are enough trees in the forest, the classifier won't overfit the model. The advantage is that the classifier of Random Forest can handle missing values.

² <https://data.worldbank.org/products/wdi>

³ <http://www.who.int/gho/en/>

4 Results

(For a more detailed information, please, consult Annex 1)

4.1 Prediction performance

It is difficult to measure the quality of a given model without quantifying its performance over training and testing. This is typically done using some type of performance metric, whether it is through calculating some type of error, the correctness of fit, or some other useful measurement. For this task, we calculate the coefficient of determination R^2 , to quantify our model's performance.

An optimal model is not necessarily a robust model. Sometimes, a model is either too complex or too simple to sufficiently generalise to new data. Sometimes, a model may use a learning algorithm that is not appropriate for the structure of the data given. At other times, the data itself could be too noisy or contain too few samples to allow a model to adequately capture the target variable, for example when the model is under fitted. R^2 and *Mean Square Error* (MSE) (see section 4.1.2 or annex 1) are the parameters that define the level of quality of our model.

4.1.1 Score

The coefficient of determination for a model is a useful statistic in regression analysis, as it often describes how "good" that model is at making predictions. The values for R^2 range from 0 to 1, which captures the percentage of squared correlation between the predicted and actual values of the target variable. A model with an R^2 of 0 always fails to predict the target variable, whereas a model with an R^2 of 1 perfectly predicts the target variable. Any value between 0 and 1 indicates what percentage of the target variable, using this model, can be explained by the features. A model can be given a negative R^2 as well, which indicates that the model is no better than one that naively predicts the mean of the target variable.

R^2 does not indicate whether the predictors are a cause of the changes in the dependent variable. We used the correlation function to measure this as well as collinearity present in the data on the explanatory variables. This measure does not show if an omitted-variable bias exists or if we have used the correct regression method. R^2 does not indicate whether the most appropriate set of independent variables has been chosen nor does it indicate whether there are enough data points to make a solid conclusion, we use the *Max-Entropy* (Steven J. Phillips, 2005) for this.

Only a subsets of the indicators used in INFORM have been entitled for the prediction. Because of the complexity of the topic, the 54 INFORM indicators are very diverse. Some types of indicators can be identified as not suitable for the proposed prediction method, namely the number of uprooted people, people affected by natural disasters, humanitarian aid. These indicators might need a dedicated modelling for predicting missing values. Other indicators did not have missing values, therefore the prediction is not required. Random Forest approach were finally found to assist the calculation of 18 indicators, of which there are 54 in total.

The following tables show of the mean R^2 score for the indicators used in INFORM (Table 2) and an example for some countries (

Table 3):

Table 2. R² of the predictions with RFR for the INFORM indicators.

Indicator Name	Indicator Id	Average R²
Agriculture Stress Index Probability	<i>ASI</i>	0.17
Corruption Perception Index	<i>CPI</i>	0.81
Net ODA received (% of GNI)	<i>ECO.DT.ODA.ODAT.GN.ZS</i>	0.93
Income Gini coefficient	<i>ECO.SI.POV.GINI</i>	0.87
Literacy rate, adult total	<i>EDU.SE.ADT.LITR.ZS</i>	0.92
Average dietary supply adequacy	<i>FS.AVA.ADSA.PR.RT</i>	0.94
Prevalence of undernourishment	<i>FS.ITK.DEFC.ZS.RT</i>	0.93
Hyogo Framework for Action	<i>HFA</i>	0.87
Children Under Weight	<i>HLT.SH.CUW</i>	0.92
Improved water source	<i>HLT.SH.H2O.SAFE.ZS</i>	0.97
Improved sanitation facilities	<i>HLT.SH.STA.ACSN</i>	0.98
Estimated number of adult living with HIV	<i>HLT.SH.DYN.AIDS.ZS</i>	0.97
Physicians density	<i>HLT.SH.MED.PHYS.ZS</i>	0.90
Maternal Mortality	<i>HLT.SH.MMR</i>	0.96
Malaria mortality rate	<i>MALARIA</i>	0.93
Human Development Index	<i>SD.HDI.UNDP.XD</i>	0.97
Gender Inequality Index	<i>SD.INEQ.GII.XD</i>	0.94
Multidimensional Poverty Index	<i>SD.MPI.UNDP.XD</i>	0.92

Table 3. Example of the R^2 score of the predictions with RFR for the INFORM indicators for some countries.

Indicator name	TUV	PRK	ATG	KNA	ERI	SOM	LBY
Agriculture Stress Index Probability	0.09	0.18	0.26	0.24	0.16	0.2	0.23
Corruption Perception Index	0.6	0.73	0.86	0.63	0.63	0.93	0.73
Net ODA received (% of GNI)	0.96	0.86	0.96	0.96	0.88	0.88	0.95
Income Gini coefficient	0.9	0.89	0.91	0.85	0.9	0.9	0.87
Literacy rate, adult total	0.83	0.92	0.89	0.89	0.94	0.84	0.89
Average dietary supply adequacy	0.91	0.95	0.95	0.95	0.97	0.93	0.97
Prevalence of undernourishment	0.9	0.91	0.86	0.84	0.97	0.96	0.96
Hyogo Framework for Action	0.83	0.79	0.88	0.9	0.84	0.84	0.87
Children Under Weight	0.95	0.92	0.83	0.83	0.96	0.91	0.89
Improved water source	0.97	1	0.98	0.98	0.9	0.98	0.95
Improved sanitation facilities	0.98	1	0.98	0.96	0.89	0.98	0.99
Estimated number of adult living with HIV	0.93	0.98	0.96	0.94	0.99	1	0.97
Physicians density	0.93	0.85	0.92	0.92	0.92	0.84	0.89
Maternal Mortality	0.97	0.99	0.9	0.93	0.9	0.97	0.99
Malaria mortality rate	1	0.93	1	1	0.93	0.85	1
Human Development Index	0.88	0.95	0.97	0.96	0.88	0.93	0.93
Gender Inequality Index	0.98	0.83	0.88	0.85	0.97	0.84	0.99
Multidimensional Poverty Index	0.92	0.9	0.88	0.89	0.97	0.89	0.91
Total average	0.84	0.85	0.86	0.84	0.84	0.84	0.87

In general countries ranges on average between 0.81 and 0.91 in R^2 , which can be considered a very positive result. Within the countries, the range of the performance of the individual indicators is quite large, varying from almost perfect to insignificant correlation.

4.1.2 MSE

The error is the measure that tells us how wrong the prediction on average. We use this method to calculate the Mean Square Error in the data set after predictions have been made. The MSE tells you how close the regression line is to a set of points. We do this by taking the distances from the points to the regression line and squaring them. The squaring is necessary to remove any negative signs. We also give more weight to larger differences. It is called the mean square error as we are finding the average of a set of errors. The smaller the MSE, the closer it is to the real value. Depending on the data, it may be impossible to get a very small value for the mean square error.

The main objective of this exercise was to have indicators in each country per year, these values would show trends that simple imputation could not show. As a result, a series of

raw data is available that replaces the old gaps in such a way that by amplifying INFORM's own methodology, a new report is created that tries to show results closer to reality. It is very important to keep in mind that predicted values are not real values: in some cases R^2 at 0 means that the predicted value is totally random and should be taken as is. This is due to non-correlation with other indicators or too much noise. In other cases the MSE could be too high or with high variance, meaning the range of predicted values result is too wide. Thanks to this new development, INFORM will have information closer to reality and will be able to present more precise predictions and trends about each of the countries in its different Social-Economic areas.

The following table show an example of the mean MSE for some countries:

Table 4. Example of the mean MSE for some countries.

Indicator name	TUV	PRK	ATG	KNA	ERI	SOM	LBY
Agriculture Stress Index Probability	0.18	0.16	0.15	0.16	0.17	0.16	0.16
Corruption Perception Index	12.99	8.53	10.49	11.4	11.57	3.99	6.72
Net ODA received (% of GNI)	1.07	1.84	0.74	0.72	2.52	1.38	1.24
Income Gini coefficient	2.46	2.58	2.35	3.81	2.4	2.46	2.84
Literacy rate, adult total	5.01	3.9	3.91	4.2	4.01	5.89	3.9
Average dietary supply adequacy	4.64	3.25	3.06	3.06	2.35	3.62	2.57
Prevalence of undernourishment	0.48	0.46	0.58	0.62	0.26	0.32	0.32
Hyogo Framework for Action	0.21	0.24	0.19	0.14	0.23	0.22	0.19
Children Under Weight	2.12	2.8	5.15	4.32	2.21	3.82	3.74
Improved water source	0.81	0.46	0.72	0.78	5.42	2.59	2.06
Improved sanitation facilities	1.69	0.84	1.88	3.07	9.12	4.01	0.99
Estimated number of adult living with HIV	1.08	0.56	1.02	1.24	0.76	0.17	0.7
Physicians density	0.3	0.47	0.32	0.32	0.35	0.58	0.5
Maternal Mortality	35.48	15.89	39.92	47.7	79.1	49.52	11.23
Malaria mortality rate	0	9.7	0	0	10.31	14.64	0
Human Development Index	0.04	0.03	0.02	0.02	0.04	0.03	0.02
Gender Inequality Index	0.03	0.05	0.05	0.06	0.03	0.04	0.02
Multidimensional Poverty Index	0.05	0.05	0.06	0.06	0.02	0.06	0.05

The following picture shows an example of predicted data. The graph demonstrates how the model transforms the information available in the data set by filling the gaps. The graph is divided into 4 graphs:

- The first one shows the original information of the data set, in blue the data is shown, the absence of blue dots indicates the lack of data in the data set.
- The second graphic in purple shows how the data set would be in the case of a direct imputation of the data, prolonging the existing values between the previous and subsequent years.
- The third graph in yellow tries to show the score value (R^2) obtained by the model for that data set, by default the existing values in the original data set are pre-assigned a value of 1 (maximum precision), the rest of the values of this graph will always have the same value because it comes from the same model.
- The fourth graph in green and red shows the prediction of values according to the model of Random Forest, in green the value itself is shown, while the red vertical lines show the MSE calculated for that model, like R^2 the MSE is unique for each model, hence the linear red has the same length.

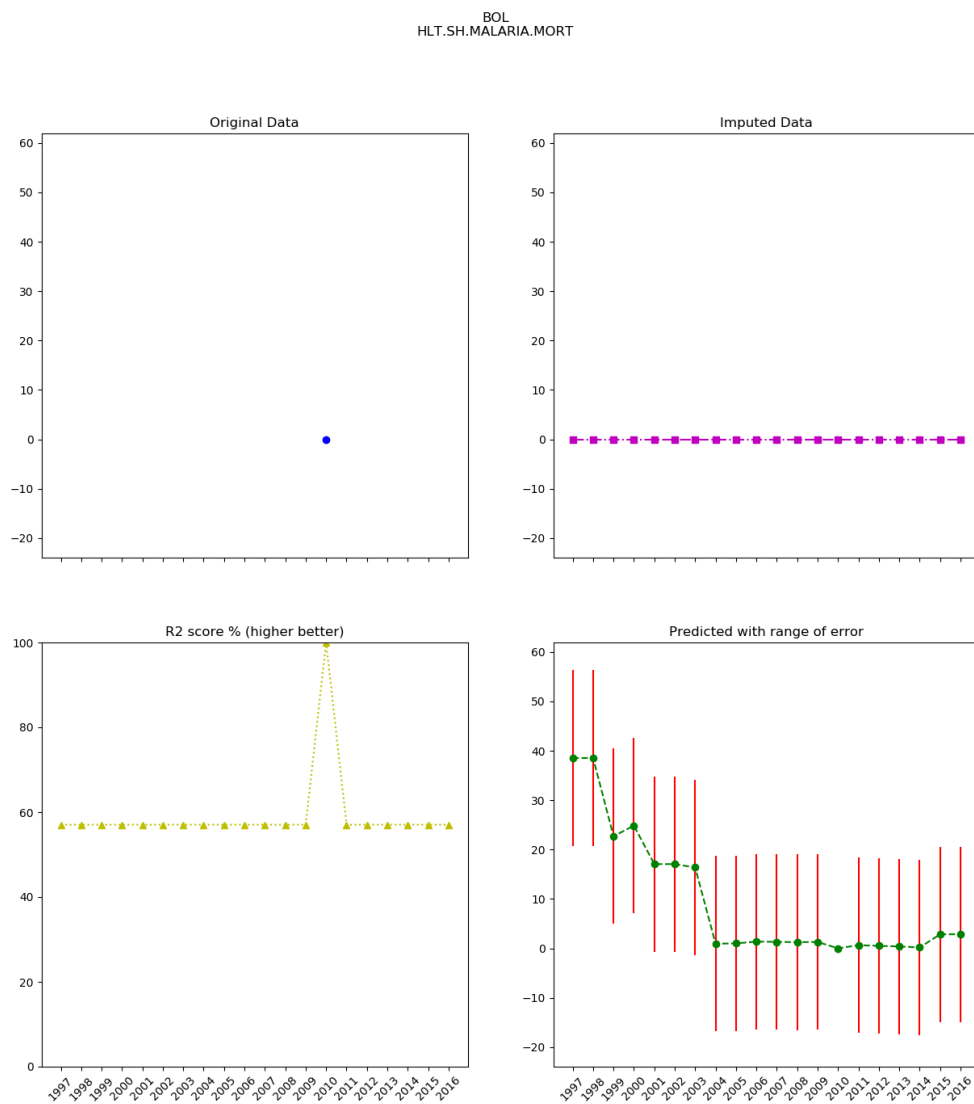


Figure 1. Predicted data for the malaria mortality rate indicator for Bolivia

4.2 Comparison with the current prediction method

Values of MSE may be used for comparative purposes. Two or more statistical models may be compared using their MSEs as a measure of how well they explain a given set of observations. In this section we compare the performances of the current INFORM prediction (see chapter 2), and the presented Random Forest method (Table 5).

Note that it was not possible to calculate the MSE for the two indicators, namely 'Prevalence of undernourishment' and 'Average dietary energy supply adequacy' (see chapter 2), for which missing values were imputed using regional average. Therefore, they are not included in the comparative analysis.

Table 5. Comparison of MSE of the RFR predictions and the current INFORM predictions for the indicators having missing values according to the last INFORM release.

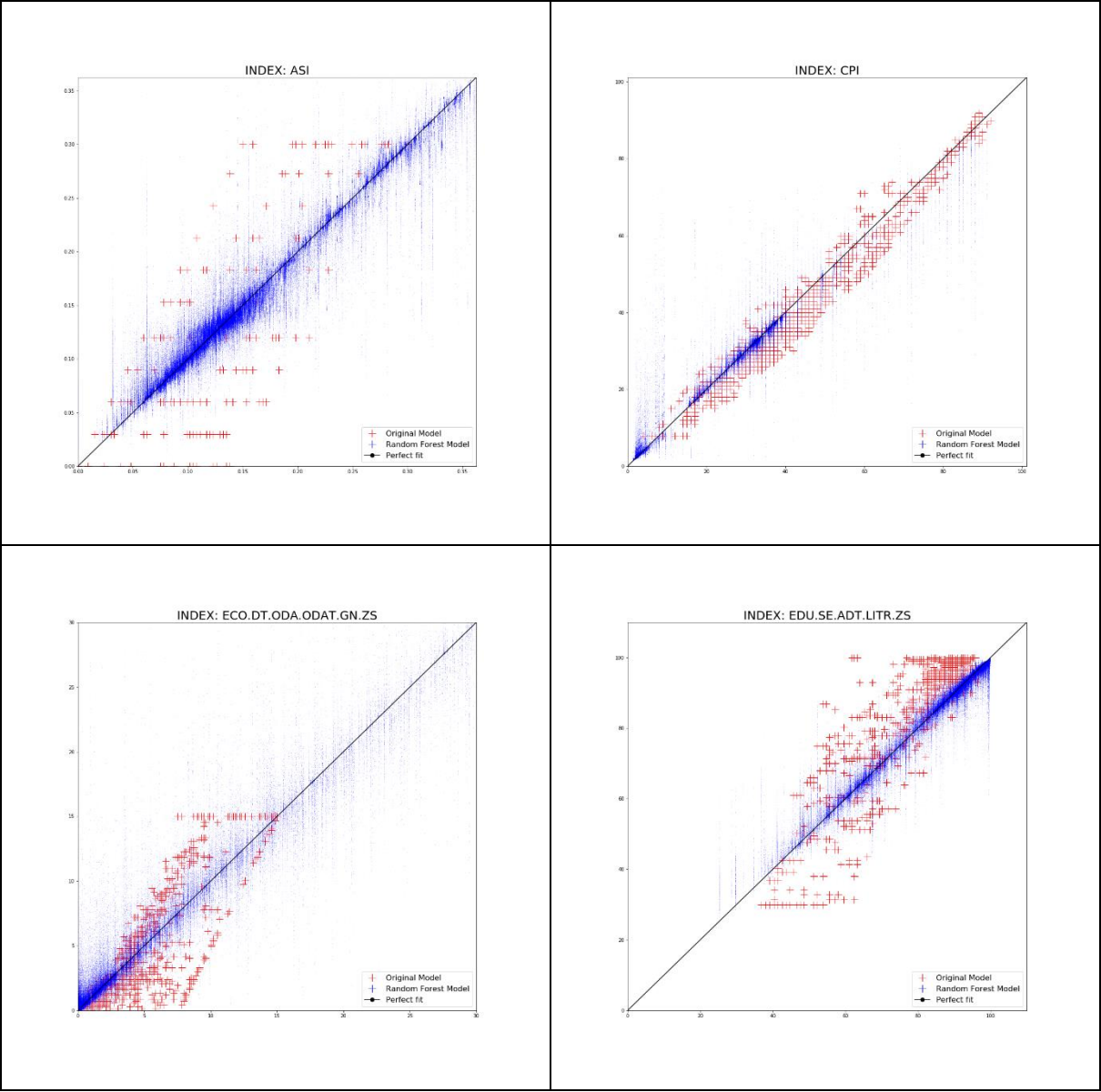
Indicator name	Average MSE [RFR]	Average MSE [INFORM]
Agriculture Stress Index Probability	0.166	0.005
Corruption Perception Index	6.592	22.143
Net ODA received (% of GNI)	1.440	5.449
Income Gini coefficient	3.032	32.278
Literacy rate, adult total	3.876	126.767
Hyogo Framework for Action	0.181	0.200
Children Under Weight	2.982	24.839
Improved water source	1.586	105.828
Improved sanitation facilities	2.588	222.505
Estimated number of adult living with HIV	0.626	0.588
Physicians density	0.362	0.906
Maternal Mortality	34.122	65413.386
Malaria mortality rate	6.973	716.882
Human Development Index	0.019	0.003
Gender Inequality Index	0.033	0.011
Multidimensional Poverty Index	0.044	0.006

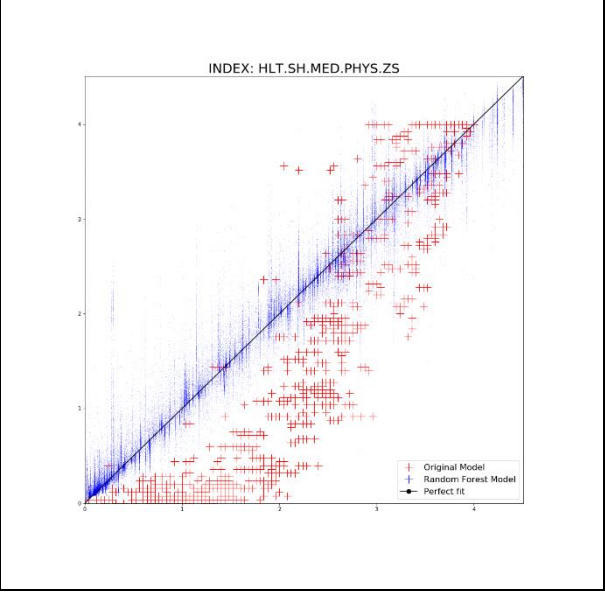
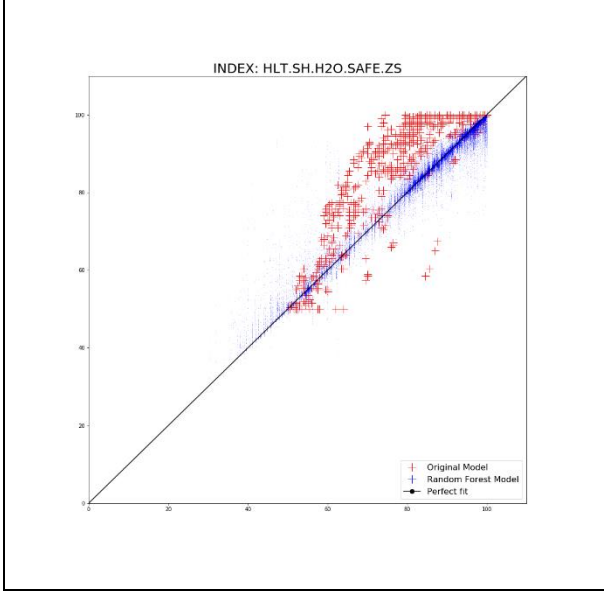
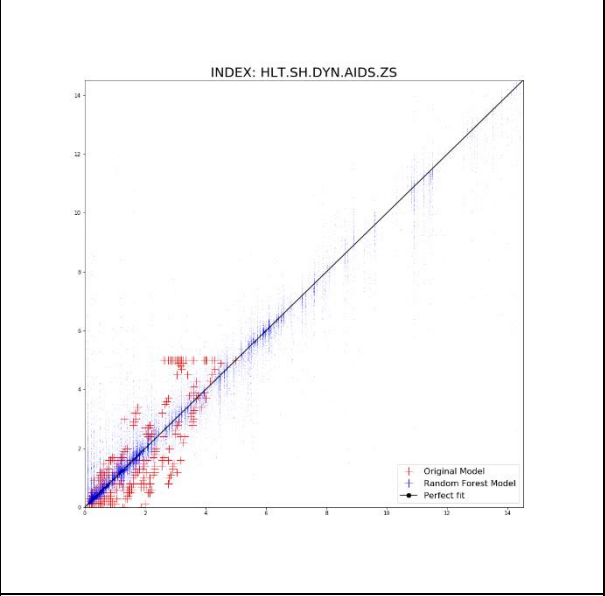
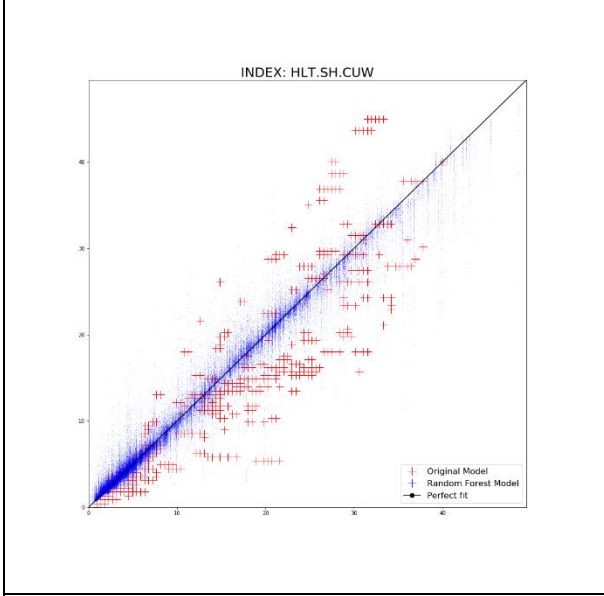
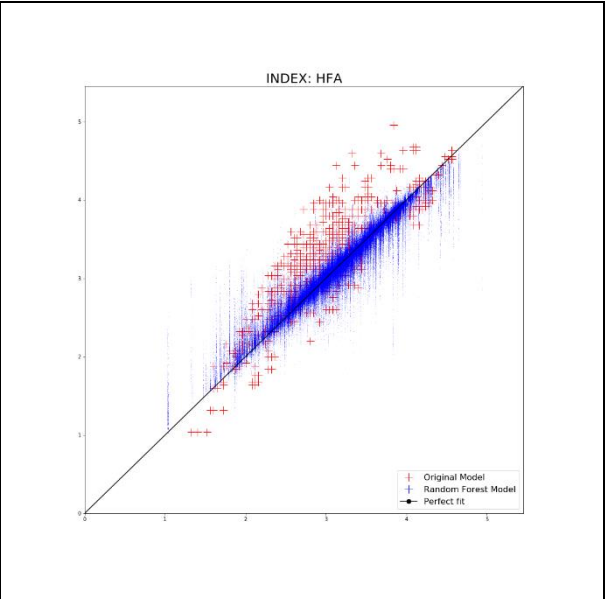
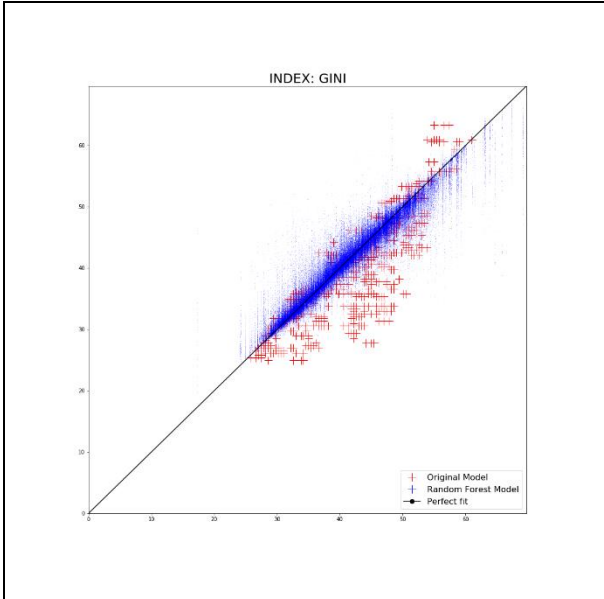
The prediction using the RFR method seems to be more efficient for most of the indicators, while for some of them (Agriculture Stress Index Probability; Human Development Index; Gender Inequality Index; Multidimensional Poverty Index) the current approach used in INFORM has still better performance.

One of the main differences between the two predictors is that the one used currently in INFORM is based only on other indicators of the same country, while the Random Forest uses all the data of all the countries. We believe that an improvement in the RFR

predictions could be achieved refining the analysis on groups of countries with similar behaviour (clusters). This might be very significant for countries with trends in contrast with the global trend (outliers), like the countries on protractive crisis e. g. conflicts).

We use real vs predicted values to also visually show the results, following plots show how the new model is most of the time closer to the perfect line (Figure 2).





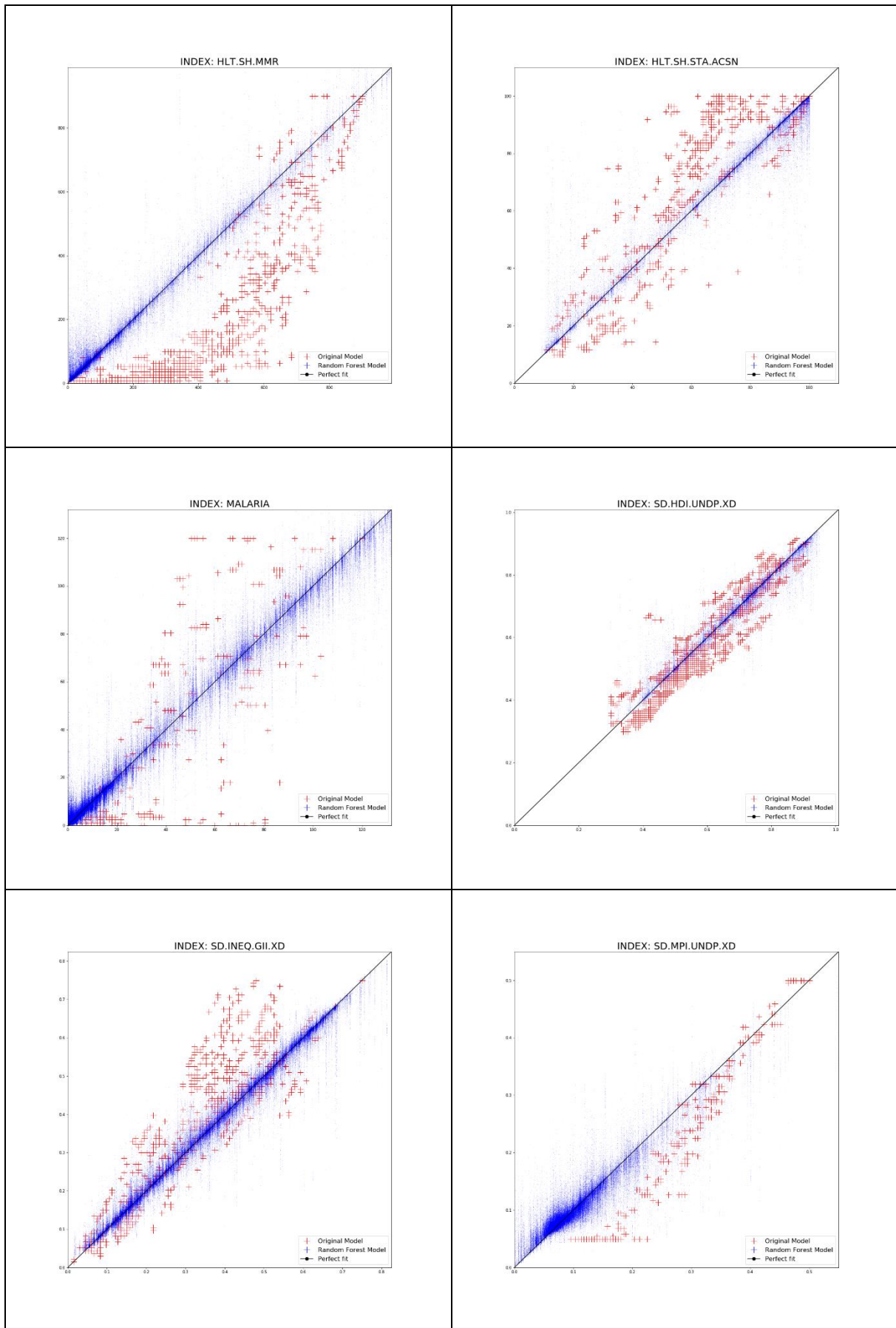


Figure 2. Real vs predicted values with RFR (blue dots), and current INFORM method (red dots).

The previous graphs (Figure 2) show the final results obtained in the new model applied to INFORM. The red dots show the predictions vs real values of the model in direct imputation showed in the current version of INFORM, the blue dots show predictions vs real values of the new model explained in this report. The black line describes the perfect fit positions, that is, the more points near the black line the better predictions the model can present. Note that the RFR method was applied to longer time series, 67 years of data, while current INFORM predictions are available for 5 years only.

4.3 Areas for Further Research

There are other techniques to apply in this predictive model not just based on the information provided by the indicator and the correlation between one and other. Unsupervised learning and Clustering is the machine learning task of inferring a function to describe a hidden structure from unlabelled data. Since the data given to the learner are unlabelled, there is no evaluation of the accuracy of the structure that is output by the relevant algorithm, however it is possible to detect hidden structures inside the data like patterns in the countries (some countries have the same behaviour in some indicators), or some indicators could be joined together in groups. This cluster model could provide new variables to inject into the current model to improve its accuracy.

Anomaly detection techniques detect anomalies in an unlabelled dataset under the assumption that the majority of the instances in the data set are normal by looking for instances that seem to fit least to the remainder of the dataset. This technique could detect patterns in the missing data providing more information about the reason o those gaps in the dataset.

Anomaly detection is applicable in a variety of domains and it is often used in pre-processing to remove anomalous data from the dataset. In supervised learning, removing the anomalous data from the dataset often results in a statistically significant increase in accuracy.

The JRC will also closely follow similar research activities promoted by INFORM partners, like the initiative of the internal displacement monitoring centre (IDMC) for predicting internally displaced people (IDPs) generating by conflicts and natural disasters, with especial interest in Climate Change related topics.

For more information, see Annex 1.

5 Data management

(For a more detailed information, please, consult Annex 2)

The JRC has developed a software tool for supporting the creation, calculation and validation of INFORM. The system supports visual analytics of results, as well as a validation workflow to guarantee quality results. The system is developed continuously according to the needs and priorities of the Inform project.

5.1 INFORM tool: external access

One objective is to prepare the INFORM calculation system for external access. Trained users can then use the central infrastructure to maintain their own INFORM-derived indexes. It is envisaged to provide this level of support to sustainable projects, such as INFORM Subnational.

The main efforts this year were focused on integrating the application for managing an INFORM model on the INFORM website. This functionality will allow users to log into the system and have dedicated access to their INFORM models. The system will allow for the full management of the INFORM models, from uploading the data to creating and updating the models.

5.1.1 Integration with DNN

DNN (DotNetNuke) is the platform used to publish the INFORM website. While the old calculation engine was an external application, the new INFORM tool has been fully integrated into the DNN platform (Figure 3).

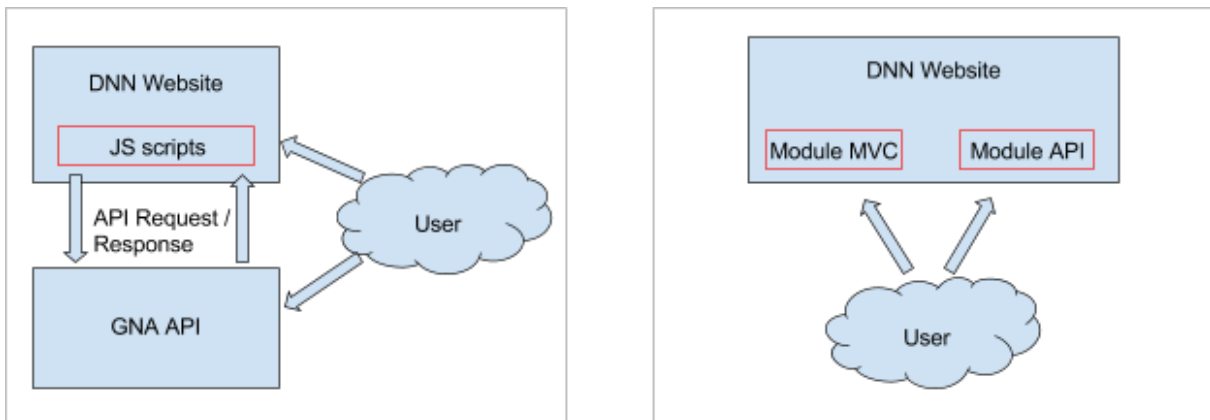




Figure 3. Differences in behaviour between old architecture (GNA app) and new Inform tool (DNN)

Users registered to the website (Figure 4 **Error! Reference source not found.**) will have different roles and will (or will not) be able to perform specific operations concerning their own role.

Visible by Administrators only.

User Management

inform

test_inform_new  

E-mail luca.vernaccini@ext.ec.europa.eu

Roles Registered Users,INFORM test,InfoRM Partner

System

inform

Workflow Groups

inform2018

Workflows

inform 2018, inform 2018 (no prediction)

Figure 4. Authentication of users and setting up of their profiles

Users may define their own INFORM Subnational model by uploading data from administrative units (shape files) for generating interactive maps (Figure 5); create a new release of their model based on existing methodology (Figure 6); or create a new methodology or modify the existing one (Figure 7).

CreateModel

If you already uploaded country data, you may [send a shape file](#) with spatial data; this let you visualize interactive maps for your Inform release.

Define name for your regional Inform model

Country Data File

No file selected.

Upload a file, or fill the box below with your data.

Note: file formats allowed are XLSX, CSV. If you upload or paste using CSV format, please specify the field separator.

Separator:

Skip first row

Figure 5. Creation of a new INFORM Subnational model

Create Workflow

New Workflow Name

INFORM GTM

Model Type Global Regional

Regional Model:

GUATEMALA

Validity - From

2017-01-01

Validity - To

2017-12-31

Workflow Group

INFORM_GTM

Or add a new one [+](#)

Methodology Template Select a methodology for calculating a composite indicator. You may copy and adapt a methodology of a previous year.

INFORM 2017 v0.3.0 (template)

Methodology Description

Description of methodology...

Comments

Use Prediction Model

Save

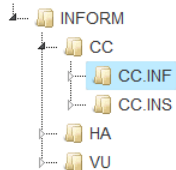
Figure 6. Creation of a new model release

Methodology Configurator

INFORM 2018 (ID: 405)

It's not possible to save any changes because the methodology is already approved. To modify the methodology, you need to [Un-Approve](#) it first.

✦ Create
✎ Rename
✕ Delete



Parent	Current Node	Children
CC	CC.INF (Infrastructure)	CC.INF.COM CC.INF.PHY CC.INF.AHC

Details

Name

Fullname

Select operator type

Select operator 🔍

Default value

VisibilityLevel

FamilyGroup

CheckPrecision

Indicators for process

Coefficient Indicator		
	CC.INF.COM	↶ 🗑️ ↷
	CC.INF.PHY	↶ 🗑️ ↷
	CC.INF.AHC	↶ 🗑️ ↷
+	✕	

CC.INF.COM;CC.INF.PHY;CC.INF.AHC;

Save

Figure 7. Configuration of the methodology

5.2 INFORM web API

A new Application Programming Interface (API) has been developed for exposing INFORM results, country profiles and all data of public interest. It will replace the old version for better performance, flexibility and variety of datasets. It will be available on the INFORM website, along with documentation and a query configurator for test purposes.

The old API will be available for a limited period of time after the release of the new version.

The guidelines for understanding how to extract data are as follows. Each INFORM model is identified by a WorkflowId. The Inform models belonging to the same release are coded with the same WorkflowGroupName (e.g. the INFORM 2017 release and the 5 years of back-calculated models based on the same methodology).

Web API Tester

Compose your search

Inform Model Type

Inform Release

Workflow

Data Type

Indicators

Geographical Area

Country

Refine search for country

Clear Filters

Active Filters:

- Global
- INFORM2017
- INFORM 2017 v0.3.1
- Results
- CC
- Western Europe
- Netherlands

Or type custom request URI: /api/

SEND REQUEST

Request URI: /API/InformAPI/countries/Scores/?WorkflowId=261&isoGroup=E2&iso3=NLD&IndicatorId=CC

```
[
  {
    i. "iso3": "NLD",
    ii. "IndicatorId": "CC",
    iii. "IndicatorScore": 1.2,
    iv. "ref_IndicatorId": null,
    v. "ref_IndicatorScore": 0,
    vi. "IndicatorRank": 0,
    vii. "Trend": null,
    viii. "FullName": "Lack of Coping Capacity Index",
    ix. "ShortDescription": "",
    x. "nodelevel": 0,
    xi. "AscDesc": null
  }
]
```

Figure 8. The INFORM API tester interface

A more technical description of the presented developments are available in the Annex 2.

6 Conclusion

All the described improvements will be presented at the INFORM Annual meeting in the mid of 2018, and then implemented in the new INFORM 2019 release.

In particular JRC will further work on the prediction of the missing data finalising to an improvement of the INFORM results and trends. The promising combination of composite indicators with machine learning tools will be further exploited with a more accurate development of the model based on clusters of countries having similar performances/behaviours.

Furthermore, JRC will complete the development of the INFORM tool, with particular focus on the user interface and the supporting user guide.

Annex 1. Predicting missing Data in INFORM

This project try to approach the moderns statistics methods to the current INFORM development. We will try to show how different mathematical methods can deal with missed data further than to apply just imputation methods from existing predictors or variables. We will be able to predict values and fill the gaps on the indexes using the information available each year and country. These predictions come with a score or error to provide a level of accuracy to the model. This project, as well as many others, use tools that take out current information, sift through data looking for patterns that are relevant to our problem, and return answers and error levels. The process of developing these kinds of tools has evolved throughout a number of fields such as chemistry, computer science, physics, and statistics and has been called machine learning, artificial intelligent, pattern recognition, data mining, predictive analytic, and knowledge discovery. While each field approaches the problem using different perspectives and tool sets, the ultimate objective is the same: to make an accurate prediction.

Introduction

The main motivations for using these new methods in INFORM, arise from the need to predict certain trends in countries that otherwise would not be possible due to the lack of information or parameters in the original data of certain countries. To achieve this, the objective is to find the maximum correlation between available information and the indicators not present.

Thanks to this new development, INFORM will have information closer to reality and will be able to present more precise predictions and trends about each of the countries in its different Social-Economic areas.

1 State of the Art

INFORM is a global, open-source risk assessment for humanitarian crises and disasters. It can support decisions about prevention, preparedness and response. INFORM is a collaboration of the Inter-Agency Standing Committee Reference Group of Risk, Early Warning and Preparedness and the European Commission.

The Inter-Agency Standing Committee (IASC) is the primary mechanism for inter-agency coordination of humanitarian assistance. It is a unique forum involving the key UN and non-UN humanitarian partners. The IASC was established in June 1992 in response to United Nations General Assembly Resolution 46/182 on the strengthening of humanitarian assistance.

1.1 About INFORM Data

The information used in INFORM come from different notoriety sources providing massive knowledge to the report, however due to its nature the raw data come with a big amount of gaps, creating some misinformation in the final results.

Current INFORM data set include information from 20 (years) x 191 (countries) x 54 (indicators), most of the information used belongs to last 5 years.

1.2 About the data

We have used data sources from World Bank Development Indicators, World Health Organization and INFORM 2017 for this research, this databases provide a wide range of predictors which let us create several different data sets and therefore different models without change the models used.

Main reason to use these sources is maximize the available information (Steven J. Phillips, 2005) relative to each country and the correlation with the values to predict. As results of this union the dataset use in this research is 67 (years) x 248 (countries) x 507 Indexes. Adding more data to the data set we are adding more information to the predicted model and improving the future score of each indicator. The main goal of adding these data is not include them in the INFORM methodology but increase the probability of having a better score in a indicator that in fact is used in the methodology.

Adding more data sometimes means add more noise, in this case Random Forest Regressor works as a filter removing that noise to a large degree.

1.3 About the Predictive models

Throughout the project we put focus in the missed indexes used in INFORM and how the models provide information about it, however these models could be applied to any variable or predictor used in the data sources.

This research is based in the field of supervised learning and the area of numerical prediction more often called regression. The other area inside supervised learning is classification out of the scope of this project.

Other field inside of machine learning is Unsupervised Learning aka. Clustering which could provide an improvement in the research that will be include in future, is out of the scope of this project.

2 About Data Analysis

Data analysis is a process of inspecting, cleansing, transforming, and modelling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, in different business, science, and social science domains. Here we show some methods used to improve the quality of the process.

2.1 Data Preparation

The data for this process comes mainly in two formats and two different sources: The first come from the Database of INFORM in MSSQL where the present information is stored without specifying where the missing data (Trevor Hastie, 2009) are. With this information a 3D array is created with axes year, country and indicator where the data not present is shown as NaN. On the other hand, another 258 new indicators are collected using the WHO and World Bank as source, these indicators also present missing data in some of their indicators. As with the initial indicators, a 3D Array with axes year, country and indicator is created. Both arrays come together to create a unique array with which the prediction process begins.

2.2 Missing Data

The method of multiple imputation (Trevor Hastie, 2009) is motivated by the need to provide an approximated value that show the current status of a given country in a given predictor. We assume that the database is analyzed by several secondary analysts, with a wide range of inferential goals and using a variety of statistical software tools and methods well suited only for complete data. We apply a small number of alternative completions, based on a model for non response. We applied the complete-data method to each completed dataset. Then results are then averaged, with an appropriate inflation for the sampling variance that reflects the uncertainty about the missing values.

2.3 Lost information

Imputation imply the loss of efficiency even that the first few imputations reduce the sampling variance substantially and latter imputations make only small contributions to the precision of the completed dataset.

The modelling and simulation steps guaranty that within and between imputation variances of the completed datasets accurately reflect the uncertainty about the missing values and it's unbiased.

2.4 Data Cleaning

Data cleaning is a common practice before to start building a model, in our case we cannot take the risk to remove more information from the dataset. Random Forest (Segal, 2004) helps in this task thanks to its good performance dealing with noise and outliers.

2.5 Reducing Predictors

Random forests are useful for feature selection in addition to being effective regressors. One approach to dimensional reduction is to generate a large and carefully constructed set of trees against a target attribute and then use each attribute's usage statistics to find the most informative subset of features. Specifically, we can generate a large set of very shallow trees, with each tree being trained on a small fraction of the total number of attributes. If an attribute is often selected as best split, it is most likely an informative feature to retain. A score calculated on the attribute usage statistics in the random forest tells us relative to the other attributes which are the most predictive attributes.

2.6 Data Exploration

Familiarizing with the data through an exploratory process is a fundamental practice to help you better understand and justify our results. Since the main goal of this project is to construct a working model, which has the capability of predicting the value of an index, we will need to separate the dataset into features and the target variables.

2.7 Type of Variables

Every predictor used in INFORM are quantitative and continuous variables hence we deal with a regression problem. Random Forest is not too sensitive to outliers or no normalized distributions, if data is not normally distributed, especially if the mean and median vary significantly (indicating a large skew), it is most often appropriate to apply a non-linear scaling, particularly for financial data. One way to achieve this scaling is by using a Box-Cox test, which calculates the best power transformation of the data that reduces skewness.

2.8 Outlier Detection

Detecting outliers in the data is extremely important in the data preprocessing step of any analysis. The presence of outliers can often skew results, which take into consideration these data points. There are many "rules of thumb" for what constitutes an outlier in a dataset. Here, we use Tukey's Method for identifying outliers: An outlier step is calculated as 1.5 times the interquartile range (IQR). A data point with a feature that is beyond an outlier step outside of the IQR for that feature is considered abnormal.

3 Developing a Model

3.1 Maximize Entropy

Entropy is the amount of information in a chosen data set. We assume the missing data do not provide information. Given a Country and a Series or indicator we have to take a sub data set where the entropy is maximum (Steven J. Phillips, 2005).

$$ME_{(C \times S)} \triangleq Cor_S \times \{\dim D_{CS} : \forall X_{ij} \notin \emptyset\}$$

Where Cor_S is the matrix correlation of S and D_{CS} is the matrix of elements not null in the subset of C x S.

3.2 Performance Metric

It is difficult to measure the quality of a given model without quantifying its performance over training and testing. This is typically done using some type of performance metric, whether it is through calculating some type of error, the goodness of fit, or some other useful measurement. For this project, we calculate the coefficient of determination, R^2 , to quantify your model's performance. The coefficient of determination for a model is a useful statistic in regression analysis, as it often describes how "good" that model is at making predictions.

The values for R^2 range from 0 to 1, which captures the percentage of squared correlation between the predicted and actual values of the target variable. A model with an R^2 of 0 always fails to predict the target variable, whereas a model with an R^2 of 1 perfectly predicts the target variable. Any value between 0 and 1 indicates what percentage of the target variable, using this model, can be explained by the features. A model can be given a negative R^2 as well, which indicates that the model is no better than one that naively predicts the mean of the target variable.

3.3 Overfitting

Overfitting is one critical problem that may make the results worse, but for Random Forest algorithm, if there are enough trees in the forest, the classifier won't overfit the model. The advantage is the classifier of Random Forest can handle missing values.

3.4 Error

The error is the measure that tell us how wrong is the prediction in its average. We use the following method to calculate the Mean Square Error in the data set after predictions have been made.

$$E_S = \frac{1}{\dim(S)} \sum_{i=\varphi_0}^{\varphi_m} (\hat{y}_i - y_i)^2$$

$$E_C = \frac{1}{\dim(C)} \sum_{k=v_0}^{v_n} E_{s_k}$$

$$= \frac{1}{\dim(S \times C)} \sum_{k=v_0}^{v_n} \sum_{i=\varphi_0}^{\varphi_m} (\hat{y}_{ik} - y_{ik})^2$$

Where m is the number of years in observations and n is the number of series, S is the vector of Series and C the vector of countries.

3.5 R² Scoring

R² is the proportion of the variance in the outputs that is predictable from the input variables, express how the hypothesis function fits the dataset.

$$\bar{y} = \frac{1}{n} \sum_{j=1}^n y_j$$

$$n\sigma^2 = \sum_{i=1}^n (y_i - \bar{y})^2$$

$$R^2 \equiv 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n\sigma^2}$$

$$R_S^2 \equiv 1 - \sum_{i=\varphi_0}^{\varphi_n} \frac{(y_i - \hat{y}_i)^2}{n\sigma^2} \quad \forall \varphi \in S$$

$$R_C^2 \equiv \frac{1}{\dim(C)} \sum_{k=v_1}^{v_n} R_{s_k}^2 \quad \forall v \in C$$

$$R_C^2 \equiv \frac{1}{\dim(C)} \sum_{k=v_1}^{v_n} \left(1 - \sum_{i=\varphi_0}^{\varphi_n} \frac{(y_{ik} - \hat{y}_{ik})^2}{n\sigma^2} \right)$$

R² does not indicate whether the predictors are a cause of the changes in the dependent variable, we used the correlation function to measure this as well as collinearity present in the data on the explanatory variables. This measure does not show if a omitted-variable bias exists or if we use the correct regression method.

R² does not indicate if the most appropriate set of independent variables has been chosen, we use the max entropy for this.

The model might be improved by using transformed versions of the existing set of independent variables but R^2 is not a indicative of this. R^2 does not show if there are enough data points to make a solid conclusion.

3.6 Shuffle and Split Data

The data is also shuffled into a random order when creating the training and testing subsets to remove any bias in the ordering of the dataset.

3.7 Training and Testing

If we don't split the data, we risk having a model that can only make good predictions with the training data set, hence, we would end up with an overfit model.

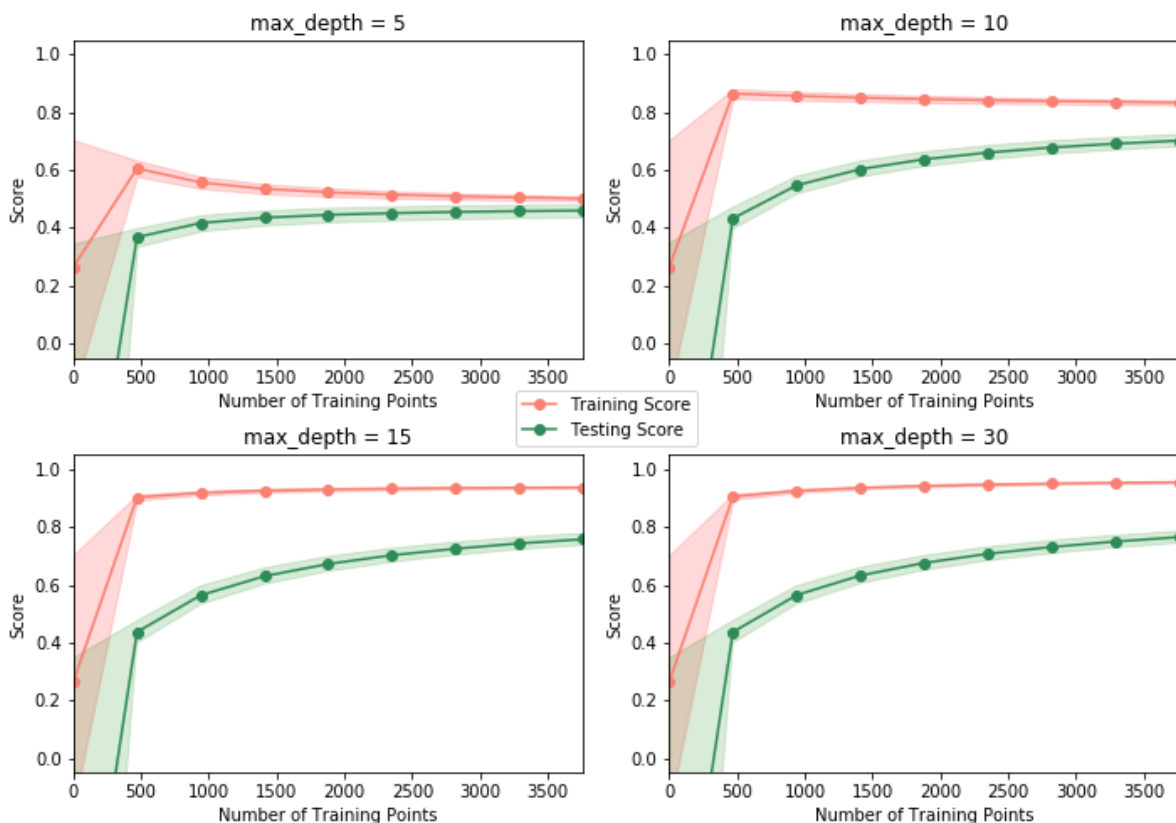
4 Analyzing Model Performance

Looking several models' learning and testing performances on various subsets of training data. Additionally, we investigate one particular algorithm with an increasing 'max depth' parameter on the full training set to observe how model complexity affects performance. Graphing the model's performance based on varying criteria is beneficial in the analysis process, such as visualizing behavior that may not have been apparent from the results alone.

A 'Random Forest Regressor' usually has a better generalization performance than an individual decision tree due to randomness that helps to decrease the model variance. Other advantages of RF are that they are less sensitive to outliers in the dataset and don't require much parameter tuning. The only parameter in RF that we typically need to experiment with is the number of trees in the ensemble and the max depth of the trees. The RF algorithm is almost identical to RF algorithm for classification, the only difference is that we use MSE criterion to grow the individual decision trees, and the predicted target variable is calculated as the average prediction over all decision trees.

4.1 Learning Curves

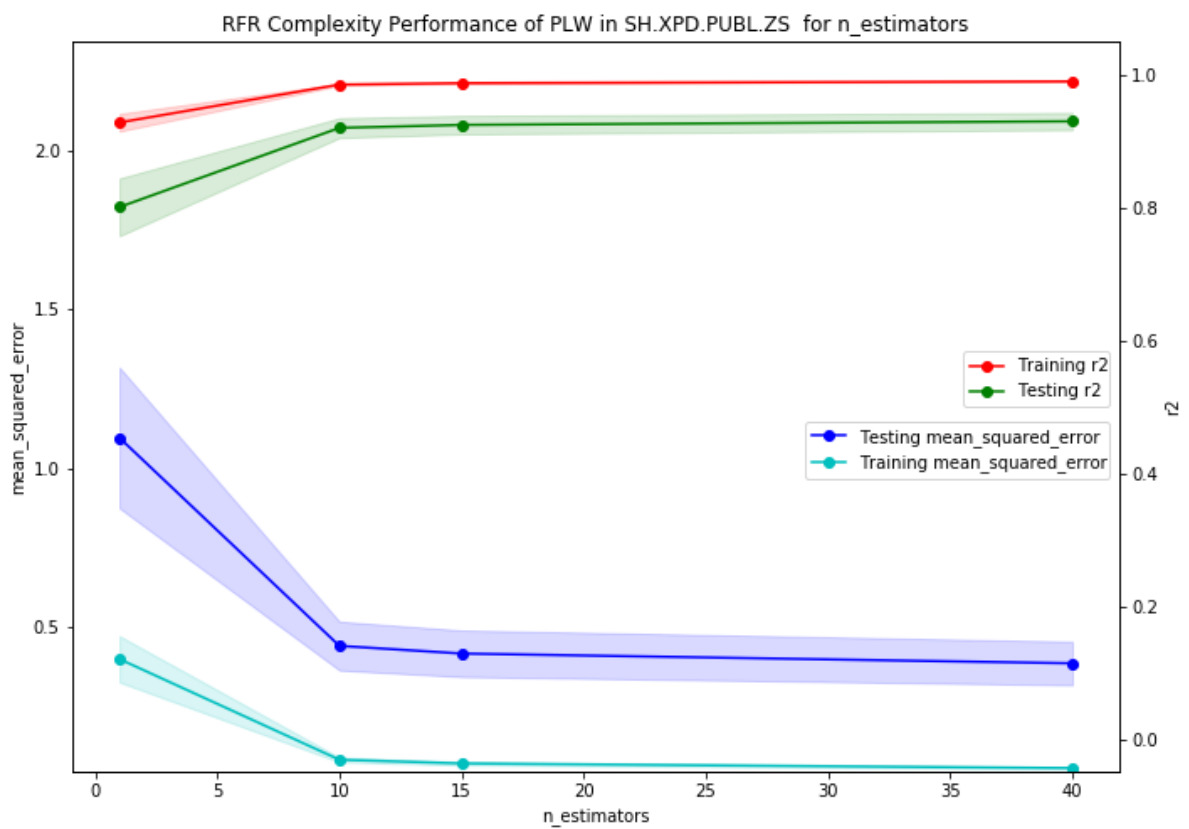
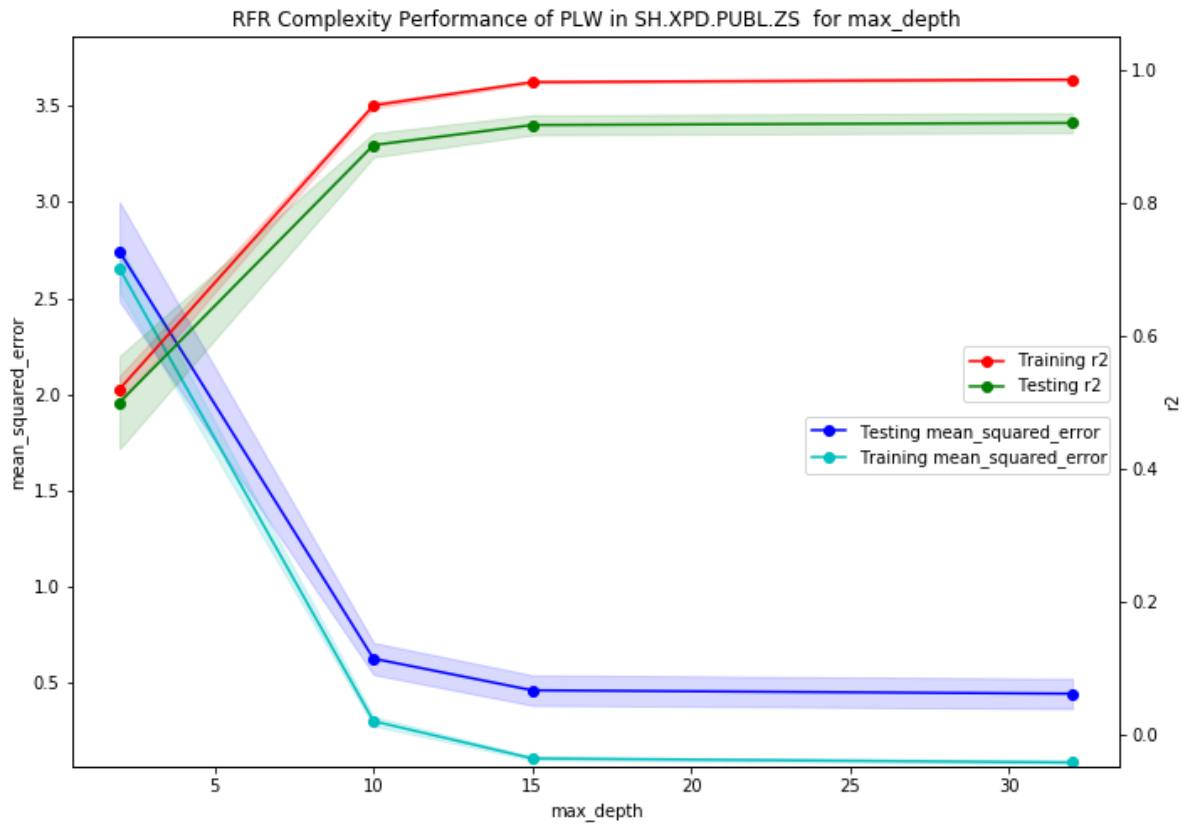
Each graph visualizes the learning curves of the model for both training and testing as the size of the training set is increased. Note that the shaded region of a learning curve denotes the uncertainty of that curve (measured as the standard deviation). The model is scored on both the training and testing sets using R^2 , the coefficient of determination.



4.2 Complexity Curves

The graph produces two complexity curves, one for training and one for validation. Similar to the learning curves, the shaded regions of both the complexity curves denote the uncertainty in those curves, and the model is scored on both the training and

validation sets using the performance metric function. The shaded regions of both the complexity curves denote the uncertainty in those curves, and the model is scored on both the training and validation sets.



4.3 Bias-Variance Trade-off

The bias-variance trade-off is a central problem in supervised learning. Ideally, one wants to choose a model that both accurately captures the regularities in its training data, but also generalizes well to unseen data. Unfortunately, it is typically impossible to do both simultaneously. High-variance learning methods may be able to represent their training set well, but are at risk of overfitting to noisy or unrepresentative training data.

In contrast, algorithms with high bias typically produce simpler models that don't tend to overfit, but may underfit their training data, failing to capture important regularities.

When the training and testing errors converge and are quite high this usually means the model is biased. No matter how much data we feed it, the model cannot represent the underlying relationship and therefore has systematic high errors.

When there is a large gap between the training and testing error this generally means the model suffers from high variance. Unlike a biased model, models that suffer from variance generally require more data to improve. We can also limit variance by simplifying the model to represent only the most important features of the data.

4.4 Best-Guess Optimal Model

In the above example, maximum depth of 15 is the Ideal Learning Curve: The ultimate goal for a model is one that has good performance that generalizes well to unseen data. In this case, both the testing and training curves converge at similar values. The smaller the gap between the training and testing sets, the better our model generalizes. The better the performance on the testing set, the better our model performs.

5 EVALUATING MODEL PERFORMANCE

5.1 Grid Search

We use *Grid Search* as a way of systematically working through multiple combinations of parameter tunes, cross-validating as it goes to determine which tune gives the best performance. The *fit* function tries all the parameter combinations, and returns a fitted classifier that's automatically tuned to the optimal parameter combination. We access the parameter values via the classifier.

A grid search algorithm guides by the performance metric and measure by cross-validation on the training set. Fine tuning a learning algorithm is a more successful learning/testing performance in terms of the application for grid search.

As we have used Random Forest Regressor the main tuning parameters that have been searched are *Max Depth* of trees (M) and *Number of trees* to use (N). Each series or indicator have their own shape, size and property which means that we cannot use a common set of parameters in the fitting.

Dealing with the issue of finding the best tuning parameters for each series, have to avoid to increase the complexity of the model giving a wide range of search in M and N . We tackled this issue giving random values to M and N in first instance. After several iteration the output of the model creates a dataset of performance metrics that we use to find the best parameters per series, given the size of the matrix.

The following table show an example about how the model works using random values in M and N , later on we will use this new dataset to build a linear regression model in charge of find the best range of M s and N s to pass these as parameters to Grid Search.

Country	Series	R^2	MSE	X	Y	Time (s)	M	N
EAR	HLT.SH.MED.PHYS.ZS	0.316	33.048	84	4	32.4	10	16
EAP	HLT.SH.TBS.INCD	0.413	14.335	32	22	11.7	12	6
DMA	IDP_B	0.000	313629.374	59	9	13.3	12	16
DOM	H.Pandemic	0.870	0.630	124	5	30.1	12	16
EAR	HLT.SH.TBS.INCD	0.424	14.198	32	22	13.0	8	6
CHN	H.Climate	0.000	2.303	61	8	19.1	13	13
CPV	H.Pandemic	0.870	0.630	124	5	31.2	12	16
CHL	H.Pandemic	0.870	0.630	124	5	33.7	12	16
DOM	IDP_B	0.000	214827.213	59	9	13.4	9	16
DZA	HLT.SH.MED.PHYS.ZS	0.000	33.108	104	20	39.9	12	17

Table 1: Sample learning dataset with Random $M \in [5, 16]$ and $N \in [5, 18]$

5.2 K-Fold Cross-Validation

Hence the K-Fold Cross-Validation (CV) estimate of prediction error:

$$CV(\hat{f}) = \frac{1}{N} \sum_{i=1}^N L(y_i, \hat{f}^{-k(i)}(x_i))$$

Where $\hat{f}^{-k(i)}(x_i)$ denotes the i^{th} fitted function with k^{th} part of the dataset removed

5.3 Fitting a Model

Our final implementation requires to bring everything together and train a model using the decision tree algorithm. To ensure that we produce an optimized model, we train the model using the grid search technique to optimize the '*max_depth*' parameter for the decision tree. The '*max_depth*' parameter can be thought of as how many questions the decision tree algorithm is allowed to ask about the data before making a prediction.

6 Making Predictions

Once a model has been trained on a given set of data, it can now be used to make predictions on new sets of input data. In the case of a Random Forest Regressor, the model has learned what the best questions to ask about the input data are, and can respond with a prediction for the target variable. We use these predictions to gain information about data where the value of the target variable is unknown, such as data the model was not trained on.

6.1 Optimal Model

As we have seen in the complexity curves, the optimal model is always linked to the complexity and some times the price we have to pay for a perfect model require infinity computational resources or time, which is not feasible in a practical environment. Find a optimal model is a never ending process that always finish under the analyst supervision, last score in this model is 88% accuracy, an improves of just 1% in the model require 1 month of computation.

6.2 Predicting Index

It's very important do not fall into temptation to get the predicted values as real, these predicted indexes are statistical values and they should be taken as it. As well as imputation that try to show values close to the real ones, but thanks to supervised learning we can provide values of how far or how accurate they are.

6.3 Sensitivity

An optimal model is not necessarily a robust model. Sometimes, a model is either too complex or too simple to sufficiently generalize to new data. Sometimes, a model could use a learning algorithm that is not appropriate for the structure of the data given. Other times, the data itself could be too noisy or contain too few samples to allow a model to adequately capture the target variable, for example the model is underfitted. R^2 and MSE are the parameters in charge to define the level of quality of our model.

6.4 Applicability

How relevant today is data that was collected from 1978?: Those data would be out of date, indexes have changed a lot during last almost 40 years and areas that in 1978 have a specific statistics, nowadays could be totally different.

Are the features present in the data sufficient to describe a missing value?: Other features should be included to predict those indexes more accuracy, for example, more sources.

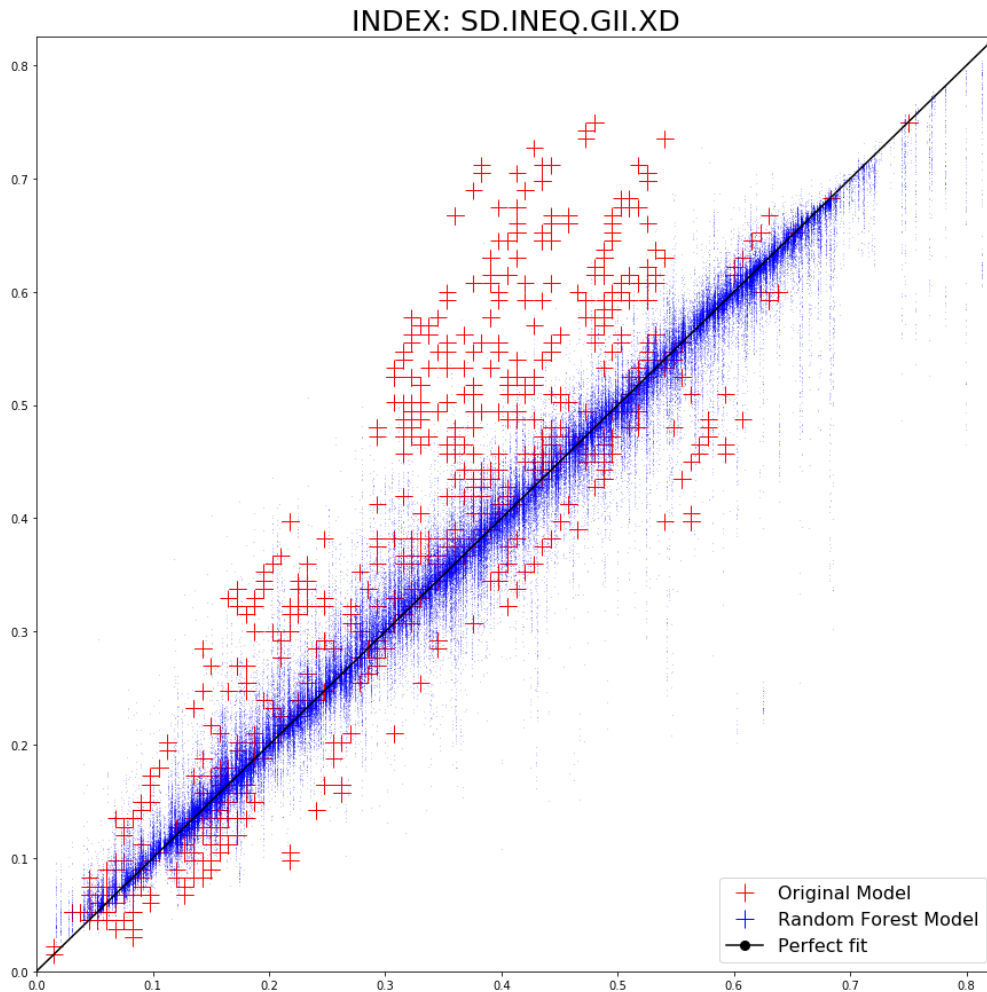
6.5 Results

The main objective of this study was to have indicators in each country per year, these values would show trends that simple imputation could not show. As a result, a series of raw data is available that replaces the old gaps in such a way that by amplifying INFORM's own methodology, a new report is created that tries to show results closer to reality. It's really important to keep in mind that predicted values are not real values, in some cases $R^2=0$ that means the predicted value is totally random and should be taken as is. This is due the none correlation with other indicators or too much noise. In other cases the Mean Square Error could be too high or with high variance therefore the range of predicted values result too wide.

Predicted indicators are included in the INFORM methodology like the real values, their don't require a special treatment, but R^2 and MSE must be included.

6.6 Model Comparison

The final goal is to compare the previous model with the new one. We use real vs predicted values to visually show the results, following plot shows an example of how the new model is closer to the perfect line.



The previous graph is an example of one of the indicators which shows the final result obtained in the new model applied to INFORM. The red dots show the predictions vs real values of the model in direct imputation showed in the current versions of INFORM, the blue dots show predictions vs real values of the new model explains in this annex. The black line describes the perfect fit positions, that is, the more points near the black line the better predictions the model can present. We can note that in the new model there is much more information having more data added to the model there that the density of blue points is much greater than the red dots.

We can conclude that the new model based on Random Forest Regressor clearly obtains better results, approximating better the predictions to the real values. As a future improvement we can observe that the representation of the *RFR* model in the previous graph shows a trend of vertical lines which indicate a clear clustering in the predictive model, which will allow better results in future versions.

Annex 2. INFORM development report

This document aims to describe the software architecture that INFORM relies on, starting from what was developed when the project was born, and moving on to changes and improvements made through years. After this overview, the focus will be on the latest modifications applied to the software architecture during 2017, including notes about further possible goals.

1. DATABASE

The RDBMS used is Microsoft SQL Server (MSSQL); the production machine is a Windows 2012 R2 Server running a 2012 SQL Server instance. The database designed for INFORM includes both tables and a programmability section with several stored procedures and functions.

1.1 Tables

COUNTRY

- Iso3 (varchar (12)): country code which, despite of the column name, can be ISO-3 or different format
- IsoGroup (varchar (12)): code for parent country
- Name (varchar (100)): name of the country
- Note (text): optional notes
- CategoryType (varchar (50)): indicates the level of depth within the country group
- CategoryInfo (varchar (100)): indicates what model the country is used for

This table collects all country data needed for the publication process, for both global and regional models. The CategoryType column can assume values like ADMIN_0[1..N] for global models, or REGION_0[1..N] for regional models.

COUNTRY INCOME

- Country (varchar (3)): Iso3 of the country
- Year (int): reference year
- Income (varchar (2)): code for income level (eg. UM = upper medium)

INDICATOR

- IndicatorId (varchar (50))
- IndicatorType (varchar (15))
- IndicatorDescription (varchar (100))
- IndicatorNote (varchar (max))
- Provider (varchar (100))
- DefaultWeight (float)
- MissingValue (float)
- Unit (varchar (10))
- IndicatorGroup (varchar (50))
- Link (varchar (255))
- Note (varchar (max))
- Copyright (varchar (max))
- Scale (varchar (50))
- Coverage (varchar (max))
- Projects (varchar (max))

This table contains definitions for the indicators used. Each methodology may, or may not use all of them, so, typically, an INFORM release is based on a subset of the indicator collection.

METHODOLOGY

- MethodologyId (int)
- WorkflowId (int)
- MethodologyDescription (text)
- MethodologyDate (datetime)
- Author (varchar (100))
- Note (text)
- Status (varchar (15))
- Iso3List (text)
- Version (varchar (50))

This table is used to store basic methodology data, like a list of countries used and the ID of the workflow it is bound to.

INDICATOR PROCESS

- [IndicatorProcessId] [int] NOT NULL,
- [ProcessId] [varchar](15) NULL,
- [IndicatorId] [varchar](50) NULL,
- [Parameters] [varchar](max) NULL,
- [GnaDefault] [float] NULL,
- [StepNumber] [int] NULL,
- [OutputIndicatorName] [varchar](50) NULL,
- [MethodologyId] [int] NULL,
- [VisibilityLevel] [int] NULL,
- [Fullname] [varchar](max) NULL,
- [Description] [text] NULL,
- [Comments] [text] NULL,
- [DataType] [int] NULL,
- [FamilyGroup] [varchar](50) NULL,
- [SortCondition] [varchar](max) NULL,
- [InfoRM_1] [text] NULL,
- [InfoRM_2] [text] NULL,
- [InfoRM_3] [text] NULL,
- [InfoRM_4] [text] NULL,
- [InfoRM_5] [text] NULL,
- [ShortDescription] [varchar](50) NULL,
- [VisibilityOrder] [int] NULL,
- [SetPrecision] [bit] NULL

This table contains all configurations needed to deploy a methodology. A detailed explanation is needed for the following columns.

- ProcessId: the operation performed to get the value for the current indicator
- IndicatorId: the name of the indicator used before the current operation
- Parameters: indicates how the current operation (ProcessId) has to be performed; depending on the process, it can be a list of inputs, threshold values or other operands
- GnaDefault: the output value to be assigned in case of null inputs
- StepNumber: the step of the process at which the current operation has to be executed. The whole process is made up of different steps, starting from 0. At the first step, the process retrieves input data from the database Every following step

is used to aggregate the indicators calculated at the previous step, until the final INFORM score is reached.

- OutputIndicatorName: the name of the indicator at the end of current operation; this name will be the IndicatorId at the next step.
- VisibilityLevel: this is a parameter used to define the visibility of the indicator into the results tables.

WORKFLOW

- [WorkflowId] [int] NOT NULL,
- [Name] [varchar](100) NULL,
- [WorkflowDate] [datetime] NULL,
- [FlagMethodologyApproved] [datetime] NULL,
- [FlagDataSaved] [datetime] NULL,
- [FlagGnaPublished] [datetime] NULL,
- [Author] [varchar](50) NULL,
- [Comments] [text] NULL,
- [GNAYear] [int] NULL,
- [System] [varchar](50) NULL,
- [WorkflowCompareId] [int] NULL,
- [GNAFromDate] [datetime] NULL,
- [GNAToDate] [datetime] NULL,
- [GNAPeriod] [bit] NULL,
- [WorkflowGroupName] [varchar](50) NULL,
- [Version] [varchar](50) NULL

This table contains all the basic data for the workflow. Even if a methodology is an abstract process and a workflow is its implementation, every time a user creates a new workflow, a new methodology row is added in the database as well, so that the relation between methodology and workflow is 1 to 1. The workflow table contains a column called WorkflowGroupName which is a sort of a tag used to identify all workflows related to the same INFORM release; this is because every INFORM release contains results for the last 5 years.

PROCESS

- [ProcessId] [varchar](15) NOT NULL,
- [ProcessType] [varchar](15) NULL,
- [ProcessDescription] [varchar](100) NULL,
- [Instruction] [text] NULL

DATAINPUT

- [ObjectId] [int] NOT NULL,
- [Iso3] [varchar](12) NULL,
- [SurveyYear] [int] NULL,
- [PubDate] [datetime] NULL,
- [InsertDate] [datetime] NULL,
- [IndicatorId] [varchar](50) NULL,
- [IndicatorValue] [float] NULL,
- [Note] [text] NULL,
- [Author] [varchar](100) NULL,
- [Source] [varchar](100) NULL,
- [GNAFromDate] [datetime] NULL,
- [GNAToDate] [datetime] NULL,
- [Version] [varchar](100) NULL

This table contains all the indicator data collected from different sources over the defined time interval. GNAFromDate and GNAToDate columns indicate time interval in which the data has to be evaluated.

DATAFINAL

- [ObjectId] [int] NOT NULL,
- [Iso3] [varchar](12) NULL,
- [SurveyYear] [int] NULL,
- [PubDate] [datetime] NULL,
- [IndicatorId] [varchar](50) NULL,
- [IndicatorScore] [float] NULL,
- [PubType] [varchar](50) NULL,
- [PubDescription] [text] NULL,
- [Note] [text] NULL,
- [Author] [varchar](100) NULL,
- [MethodologyId] [int] NULL,
- [GNAYear] [int] NULL,
- [GNAFromDate] [datetime] NULL,
- [GNAToDate] [datetime] NULL,
- [Version] [varchar](100) NULL,
- [OidDatainput] [int] NULL

This contains scores for all methodologies deployed.

OPTIONS

- [TableName] [nvarchar](50) NOT NULL,
- [FieldName] [nvarchar](50) NOT NULL,
- [ID] [int] NOT NULL,
- [Value] [nvarchar](50) NULL

This table is used to store configurations related to methodologies, such as the number of decimal points to be used in the results, whether the model is regional or not, and so on.

OPTIONSCOMBO

- [TableName] [nvarchar](50) NOT NULL,
- [FieldName] [nvarchar](50) NOT NULL,
- [ID_Combo] [int] NOT NULL,
- [DES_Combo] [nvarchar](50) NULL

This contains the list of the models available.

Note: there are other tables in the database which are not used anymore, or that have been created to be of use for subtasks like data import which are separate concerns and will be investigated later in this document.

1.2 Programmability - Stored Procedures

The database currently contains a long list of stored procedures written during a second phase of development with the aim of removing all SQL queries originally injected in the source code.

This document reports on details for only a few of them; those considered the most important and/or complex.

```

CREATE Proc [dbo].[usp_Indicators]
    @WorkflowId INT = 0,
    @MethodologyId INT = 0,
    @StepNumber INT=-1,
    @MaxVisibility INT=-1
AS
BEGIN

    declare @MaxVisibilityGeneral int

    if(@WorkflowId > 0)
        begin
            select @MethodologyId = MethodologyId
            from Methodology
            where WorkflowId = @WorkflowId
        end

    select @MaxVisibilityGeneral = max(VisibilityLevel)
    from IndicatorProcess p
    where MethodologyId = @MethodologyId

    select p.*
    from IndicatorProcess p
    where p.MethodologyId = @MethodologyId
    and StepNumber = case when @StepNumber > -1 then @StepNumber else
StepNumber end
    and VisibilityLevel <= case when @MaxVisibility > -1 then @MaxVisibility else
@MaxVisibilityGeneral end
    order by StepNumber desc, OutputIndicatorName

END
GO

```

This procedure retrieves all Indicator Process by WorkflowId or MethodologyId.

```

CREATE PROCEDURE [dbo].[usp_GetDataAvailability2]
    @WorkflowId int,
    @IndicatorIdPar varchar(50) = null,
    @Iso3Par varchar(max) = null,
    @UsePrediction tinyint

as
begin
    declare
        @IndicatorId varchar(50),
        @OutputIndicatorName varchar(50),
        @SelectMethod varchar(50),
        @Iso3List varchar(max),
        @dateFrom DateTime,
        @dateTo DateTime,
        @Parameters varchar(100),
        @Version varchar(100),
        @rc cursor

    declare @TMP_IND table (id int, IndicatorId varchar(max))

```

```

insert into @TMP_IND select id, [Data] as IndicatorId from ufn_Split(@IndicatorIdPar,
',')

create table #TempData (Iso3 varchar(12), SurveyYear int, IndicatorId varchar(50),
IndicatorValue float, r2 float, mse float, Author varchar(100), Source varchar(100),
PubDate DateTime, InsertDate DateTime, FromDate DateTime, Note text, ToDate
DateTime, Version varchar(100), IsPredicted tinyint, CountryName varchar(100),
IndicatorDescription varchar(255))

set @rc = cursor for
select
case when SUBSTRING(Parameters, (PATINDEX('%VERSION=%', Parameters)+8),
LEN(Parameters) - PATINDEX('%VERSION=%', Parameters)) <> '' then
SUBSTRING(Parameters, (PATINDEX('%VERSION=%', Parameters)+8), LEN(Parameters)
- PATINDEX('%VERSION=%', Parameters)) else null end as [Version],
IndicatorId,
OutputIndicatorName,
right(ProcessId, 3) as SelectMethod,
case when @Iso3Par is null then replace(convert(varchar(max), Iso3List), ';', ',')
else @Iso3Par end as Iso3List,
DATEADD(day, convert(int, substring(Parameters, PatIndex('%[0-9,-]%',
Parameters), PATINDEX('%;%', Parameters) - PatIndex('%[0-9,-]%', Parameters))),
w.GNAFromDate) as dateFrom,
DATEADD(day, convert(int, substring(Parameters, (PATINDEX('%TO=[0-9,-]%',
Parameters) +3), PATINDEX('%;VERSION%', Parameters) - (PATINDEX('%TO=[0-9,-]%',
Parameters) +3))), w.GNAToDate) as dateTo
from IndicatorProcess ip
join Methodology m on m.MethodologyId = ip.MethodologyId
join WorkFlow w on w.WorkflowId = m.WorkflowId
where m.WorkflowId = @WorkflowId
and ((@IndicatorIdPar is null) or (IndicatorId in (select IndicatorId from @TMP_IND)))
and StepNumber = 0

open @rc
fetch next
from @rc into @Version, @IndicatorId, @OutputIndicatorName, @SelectMethod,
@Iso3List, @dateFrom, @dateTo
while @@FETCH_STATUS = 0
begin
insert into #TempData exec [usp_GetFromDataInput2] @IndicatorId,
@OutputIndicatorName, @SelectMethod, @Iso3List, @dateFrom, @dateTo, 0, null, 0,
@UsePrediction
fetch next
from @rc into @Version, @IndicatorId, @OutputIndicatorName, @SelectMethod,
@Iso3List, @dateFrom, @dateTo
end

close @rc
deallocate @rc

select * from #TempData

drop table #TempData

end
GO

```

This procedure retrieves all DataInput by WorkflowId; the complexity here can be explained by the need to identify the correct input for each process of the methodology, in terms of time interval and version of the data.

```

CREATE procedure [dbo].[usp_GetFromDataInput2]
    @IndicatorId varchar(50),
    @OutputIndicatorName varchar(50) = null,
    @SelectMethod varchar(50) = null,
    @Iso3List varchar(max) = null,
    @dateFrom DateTime = null,
    @dateTo DateTime = null,
    @SurveyYear int = 0,
    @Version varchar(50) = null,
    @YearRef int = 0,
    @UsePrediction tinyint,
    @IsGlobal tinyint = 1
as
begin

declare @TMP_ISO3 table (id int, Iso3 varchar(max))
if @Iso3List is not null
    begin
        insert into @TMP_ISO3 select id, [Data] as Iso3 from ufn_Split(@Iso3List, ',')
    end
else
    begin
        insert into @TMP_ISO3 select Iso3 as id, Iso3 from Country where CategoryType =
'ADMINO'
    end

if @SelectMethod is null
    select i.Iso3, @OutputIndicatorName as IndicatorId, SurveyYear, InsertDate, PubDate,
IndicatorValue, i.Note, Author, GNAFromDate as FromDate, GNAToDate as ToDate,
Version, IsPredicted, CountryName, IndicatorDescription
    from DataInput i
    join Country c on c.Iso3 = i.Iso3 and substring(c.CategoryInfo, 1, 6) = case when
@IsGlobal = 1 then 'INFORM' else 'REGION' end
    join Indicator ind on ind.IndicatorId = i.IndicatorId
    where i.IndicatorId = case when @IndicatorId is null then i.IndicatorId else
@IndicatorId end
    and SurveyYear = case when @SurveyYear = 0 then SurveyYear else
@SurveyYear end
    and IsPredicted = case when @UsePrediction = 1 then IsPredicted else 0 end
    and i.Iso3 in (select Iso3 from @TMP_ISO3)
    and IndicatorValue <> -99
    and (([Version] is null and @Version is null) or ([Version] = " and @Version is null) or
([Version] = @Version and @Version is not null) or (@Version is null and [Version] =
[Version] and @UsePrediction = 1 and IsPredicted = 1))
    and GNAToDate >= case when @dateFrom is null then GNAToDate else @dateFrom
end
    and GNAToDate <= case when @dateTo is null then GNAToDate else @dateTo end
    order by Iso3, GNAToDate

else
    begin
        select i.Iso3, max(SurveyYear) as SurveyYear, @OutputIndicatorName as

```

```

IndicatorId, IndicatorValue, min(i.r2) r2, min(i.mse) mse, Author, [Source], i.PubDate,
i.InsertDate, i.GNAFromDate as FromDate, min(convert(varchar(max), i.Note)) as Note,
i.GNAToDate as ToDate, [Version], IsPredicted, CountryName, ind.IndicatorDescription
from DataInput i
join
(
select i3.Iso3, i3.IndicatorId, i3.GNAToDate, max(i3.PubDate) as PubDate
from DataInput i3
join (
select Iso3, IndicatorId, max(GNAToDate) as GNAToDate
from DataInput
where IndicatorId = @IndicatorId
and GNAToDate >= @dateFrom
and GNAToDate <= @dateTo
and IndicatorValue <> -99
and (([Version] is null and @Version is null) or ([Version] = "" and @Version is
null) or ([Version] = @Version and @Version is not null) or (@Version is null and
[Version] = [Version] and @UsePrediction = 1 and IsPredicted = 1))
and IsPredicted = case when @UsePrediction = 1 then IsPredicted else 0 end
group by Iso3, IndicatorId
) i4 on i4.Iso3 = i3.Iso3 and i4.IndicatorId = i3.IndicatorId and i4.GNAToDate
= i3.GNAToDate
where IndicatorValue <> -99
and i3.GNAToDate >= @dateFrom
and i3.GNAToDate <= @dateTo
and (([Version] is null and @Version is null) or ([Version] = "" and @Version is
null) or ([Version] = @Version and @Version is not null) or (@Version is null and
[Version] = [Version] and @UsePrediction = 1 and IsPredicted = 1))
and IsPredicted = case when @UsePrediction = 1 then IsPredicted else 0 end
group by i3.Iso3, i3.IndicatorId, i3.GNAToDate
) i2 on i2.Iso3 = i.Iso3 and i2.IndicatorId = i.IndicatorId and i2.PubDate =
i.PubDate and i2.GNAToDate = i.GNAToDate
join Country c on c.Iso3 = i.Iso3 and substring(c.CategoryInfo, 1, 6) = case when
@IsGlobal = 1 then 'INFORM' else 'REGION' end
join Indicator ind on ind.IndicatorId = i.IndicatorId
where i.Iso3 in (select Iso3 from @TMP_ISO3)
group by i.Iso3, i.IndicatorId, i.IndicatorValue, i.InsertDate, i.GNAFromDate,
i.GNAToDate, i.Author, Source, i.PubDate, [Version], IsPredicted, CountryName,
ind.IndicatorDescription
end
end

GO

```

This procedure implements the logic for data extraction, based on several parameters. The first filter is the UsePrediction flag, which specifies whether we want to look up both real and imputed data, or real data only.

Since the database stores different versions of the same indicator, the query looks for the greatest "GNAToDate" which represents the final validity of the indicator and, if duplicates are found, selects the indicator with latest "PubDate" (publication date).

2. .NET SOLUTION

INFORM is based on a .NET Web Application. Let's see how the first version, called GNA, was developed.

2.1 Architecture

The GNA solution is a **Web Forms** Application used by admin users to load data, manage methodologies, run workflows and publish results.

The main solution contains three different projects:

GNA_Connector

This is a collection of classes used for different purposes, such as a data access layer, business logic and simple datasets representation.

Here is an overview of the main classes referenced.

- GNASqlDb. This works as a data access layer and it contains methods to handle connections and transactions; also, all stored procedures are called from this class, which still includes a series of queries directly in the code, hence the high number of lines.
- GNAHelper. This is a static class that contains most of the business logic; it stores an instance of GNASqlDb as a static member, in order to use a static connection to the database (DB).
- GNACalculation. This contains the implementation of every process type.
- GNAWorkflow. This is the Workflow model that can be bound to the DB table.

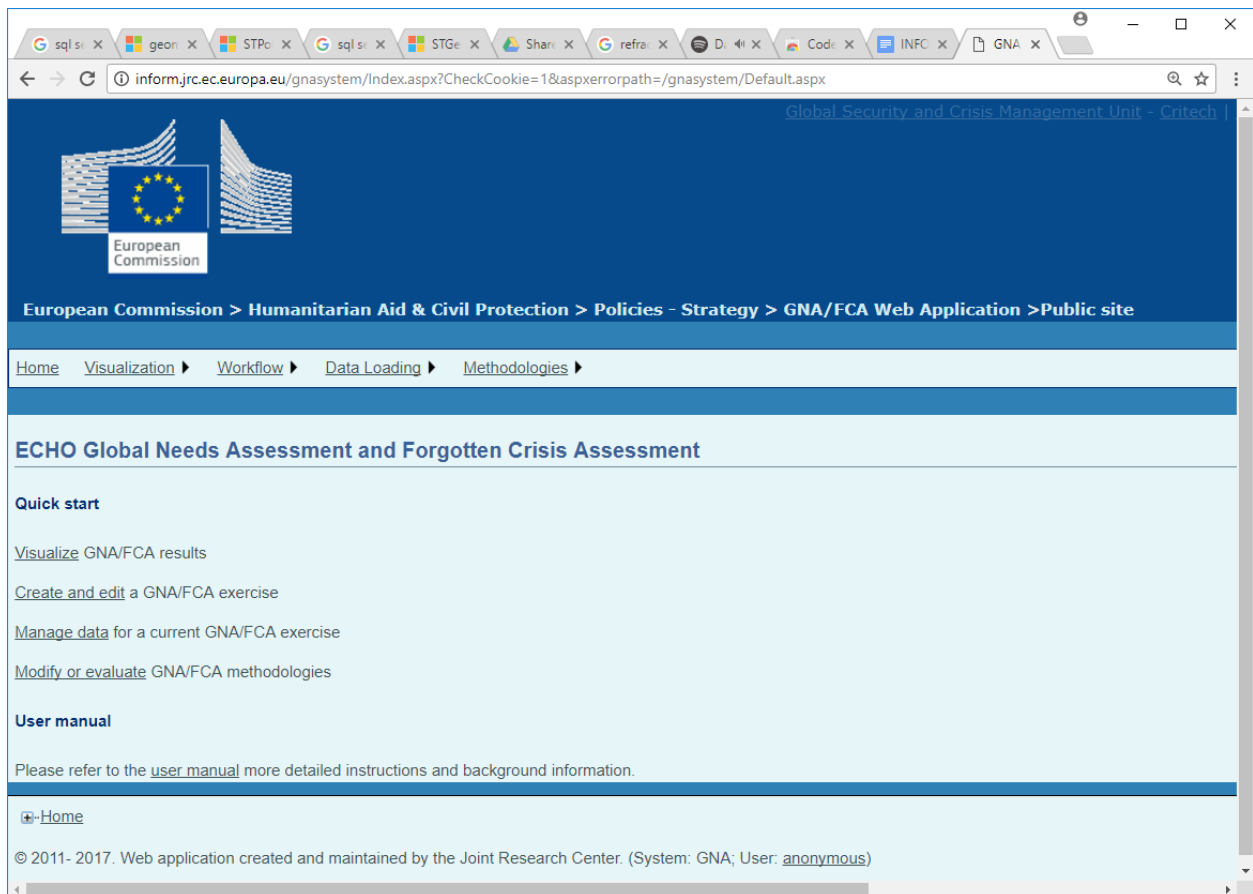
GNA_Uilities

This only contains a static class Utilities which contains generic utility methods, like a shared helper.

GNA_Webapplication

This is the event driven Web Forms application. Here follows a short explanation of how it works.

The original application was published on the <http://inform.jrc.ec.europa.eu> website and it featured the main functionalities to create a methodology and publish its results.



- *Figure 2.1. GNA web application*

2.2 Web API

Having INFORM results only visible on a webpage didn't meet third parties' need to easily access the data, so the application grew with a web API meant to display results to the public.

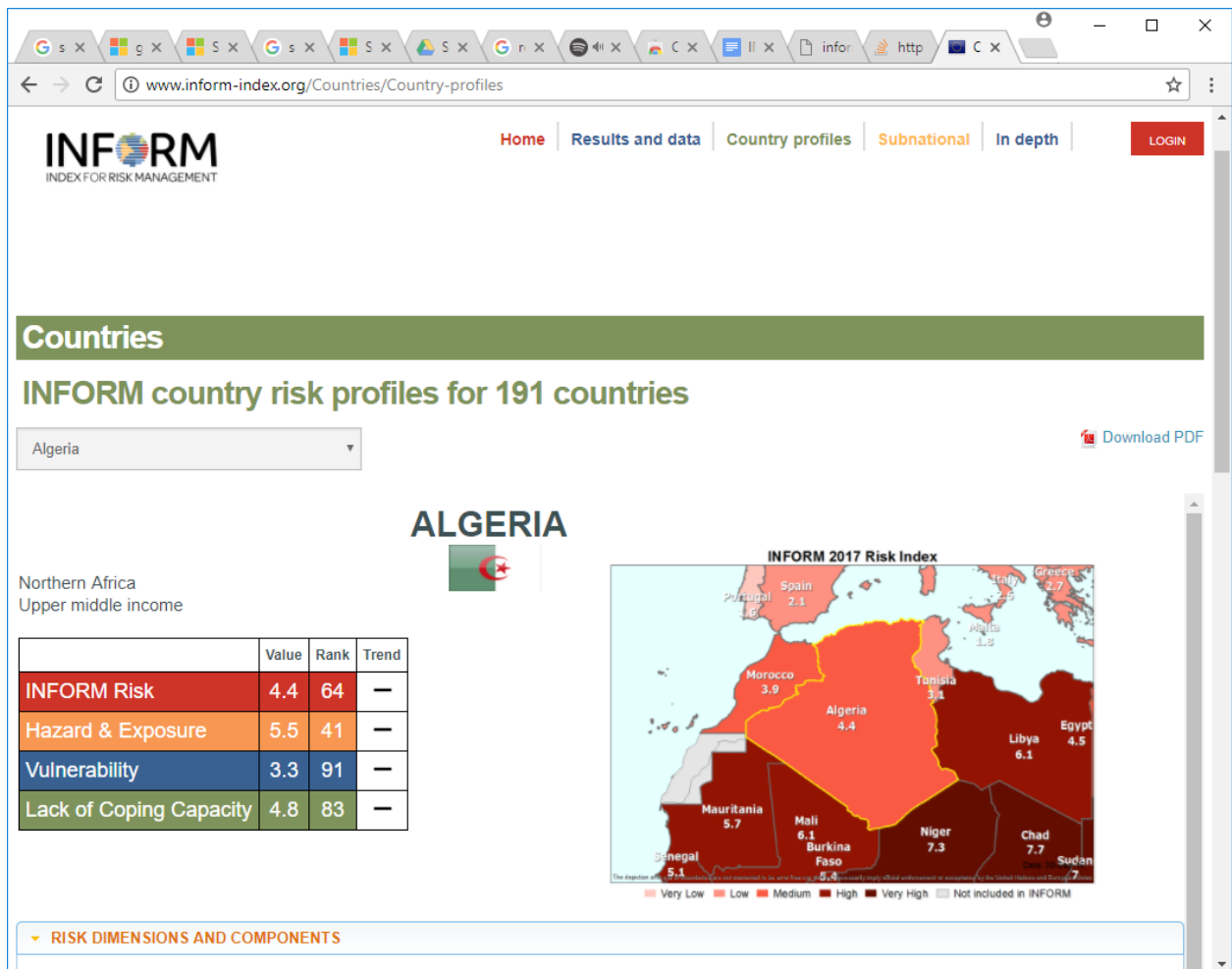
This web API was actually a workaround, because the architecture of the GNA_Webapplication didn't allow for the creation of controllers or manage routing like a modern RESTful API. So, basically, a page was created called `api001.aspx` that contained a collection of methods used to retrieve a number of different datasets based on INFORM results.

A list of available API calls is visible at this address:

http://inform.jrc.ec.europa.eu/gnasystem/APIDocumentation/API_documentation.html

When the INFORM website was published on a DNN (DotNetNuke) platform (<http://www.inform-index.org/>), the problem was about how to show the results on this website, having the core application deployed on another machine and responding to a different URL.

The solution, at first, consisted of using iFrames to include pages from the GNA website. An example would be the Country Profile section on the INFORM website:



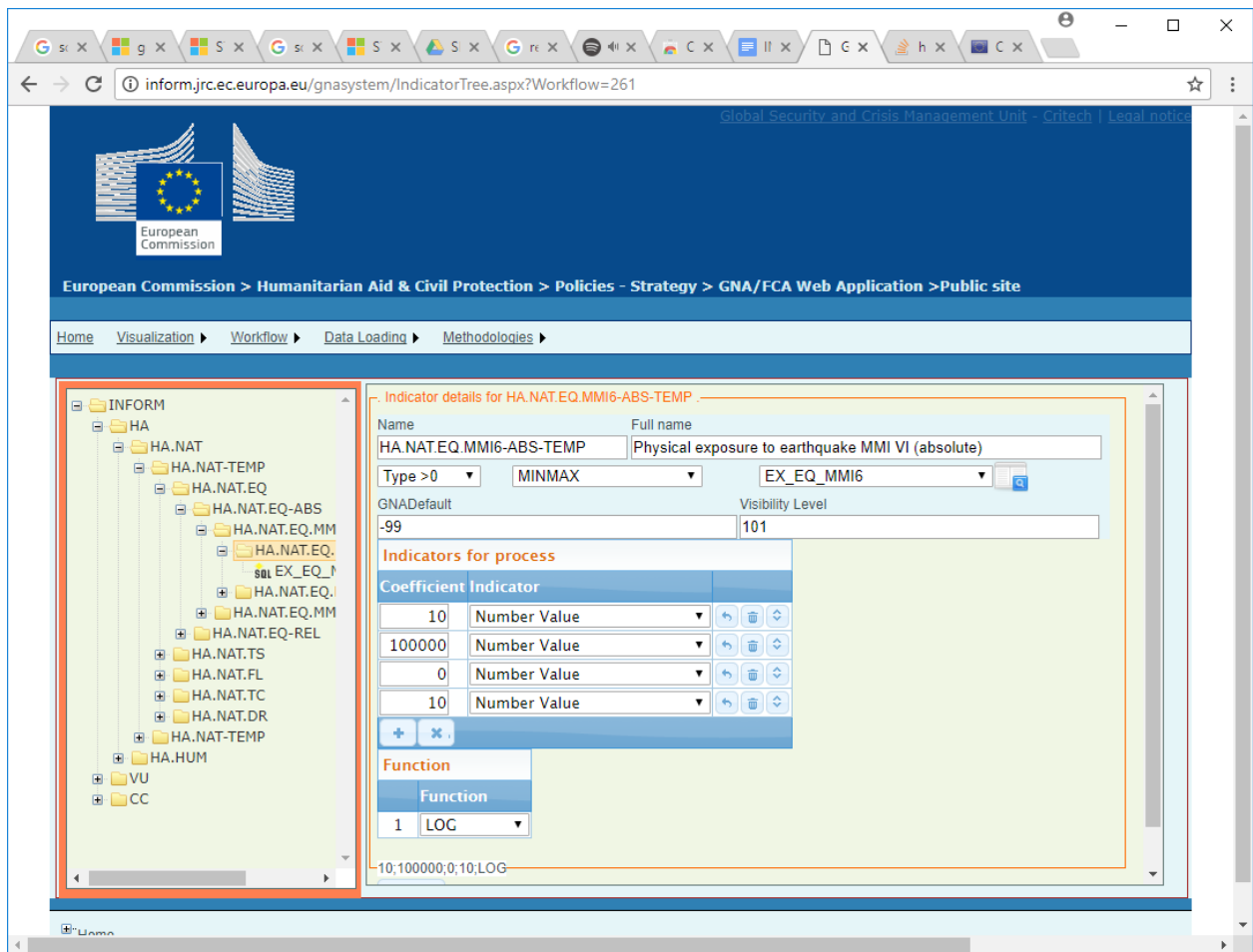
- *Figure 2.2. INFORM country profile*

In this example, the URL

http://139.191.244.117/gnasystem/isochoice_iframe.aspx?iso3=DZA&workflow=261&workflowgroup=INFORM2017 is included in <http://www.inform-index.org/Countries/Country-profiles> by using iFrame.

The second part of the solution was to extend the web API to make more of the core functionalities available remotely.

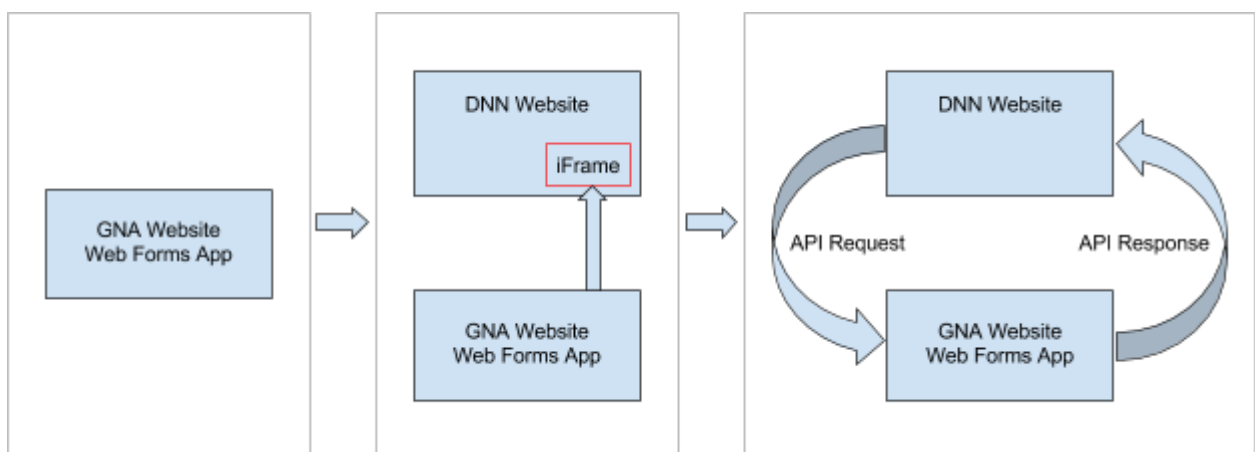
An example would be the methodology configurator called the Indicator Tree whose dynamic layout is built with javascript and interacts with the server via ajax calls.



• *Figure 2.3. Methodology configurator*

This screenshot shows details of the configuration for the methodology linked to workflow n.261

So, a summary of how the application has changed during the years would be:



• *Figure 2.4. Evolution of web architecture*

By the end of 2016 the development of the API and the integration with DNN was not fully completed. Here is a list of the functionalities already implemented at that time

- Methodology creation => only on old GNA website
- Methodology configuration *
- Methodology approval & publishing => only on old GNA website
- Country Profile *
- Interactive map (for INFORM Subnational models) *
- Web API for exposing results
- Data upload with Excel files => only on old GNA website
- Data import from external sources: World Bank, World Health Organisation (WHO), Office of the United Nations High Commissioner for Refugees (UNHCR) => only on old GNA website

*data load by Javascript

2.3 Other functionalities

The GNA app also included functionalities to import data from external sources; the connections implemented so far are with the World Bank, WHO and Human Development Report (HDR) databases. A separate project was built to connect to UNHCR APIs because of its particular requirements.

The database contains several tables created to store configurations regarding these processes. The tables are as follows:

API_IND_Conversion
 API_IND_Conversion_Rules
 API_IND_Conversion_RulesAttributes
 API_INDICATORS
 API_Region
 API_Region_Attributes
 API_Region_Country
 API_Rules
 API_Rules_Attributes
 API_UNHCR_DATA_POPDATA
 API_UNHCR_POPDATA
 API_UNHCR_Settlements

3. What's next?

Starting with the second half of 2017, the development of INFORM applications was resumed. This chapter explains the results of the software analysis.

3.1 Issues with old application

First of all, it was clear that the old GNA application was hindered by several issues:

Old architecture

The architecture of the application was a little outdated and it no longer met requirements in terms of flexibility and performance. Web Forms may still be suitable for small applications, or at least for applications with limited user interaction, but they are not suited to the most recent needs, as they are not conceived for implementing web services.

Slowness

Most of the operations took very long to be executed, and this was not acceptable considering the need to make INFORM available to a wider audience.

Lack of modularity

The continuous development made by different programmers over the years without a modular approach, made the code over-complex, redundant and difficult to maintain.

Lack of abstraction

The code was full of queries written to retrieve many specific datasets which were mapped on dedicated classes, which made the code very complex and difficult to extend.

3.2 Solution proposed

So, how to solve these problems?

The best approach identified consists in taking only the calculation engine from the old application and building a completely new one, using a model view controller (MVC) pattern.

Explaining what MVC is and what architectural patterns are is not what this document is intended for, so we will simply focus on the benefits expected from the solution proposed.

1. Enables full control over the rendered HTML.
2. Provides clean separation of concerns (SoC).
3. Enables Test-Driven Development (TDD).
4. Provides easy integration with JavaScript frameworks.
5. Follows the design of the stateless nature of the web.
6. Uses RESTful URLs that enable SEO.
7. Generates no ViewState orPostBack events.

If the above list is nothing but an understood and agreed comparison between MVC and WebForms, it is clear that those differences could provide us with a more scalable and robust application.

3.3 INFORM Tool Solution

INFORM Tool Solution is the working name given to this new project that should replace the old GNA System.

Let's see what has been done so far.

3.3.1 Architecture

As mentioned above, the new application uses an MVC pattern, but in the end, it will be a mixture of MVC and Web API, because of the need to have a public API to expose INFORM results.

Models and object relational mapping

The GNA System did not use an object relational mapping (ORM), but it just executed SQL queries from the code, mapping the results on ad hoc model classes without any abstraction; so every time there was the need to handle even a slightly different dataset, a new class would be created, or arrays would be used to handle data. During the second part of GNA System development, most of the queries were moved from the code to the database server as stored procedures, but the lack of abstraction issue was still there. The new application tries to simplify things in two ways:

1. by identifying stored procedures that return similar datasets and replacing them with more generic ones;
2. by identifying classes that refer to the same abstract entity and replacing them with a new one which has the union of members.

A further action needed to eliminate redundancy is the creation of abstract classes, leveraging inheritance and override features.

This way the application is still not using an ORM like Entity Framework or LinqToSql, which would certainly provide a simpler code to handle DB objects and easier development of the public API as a result. However, the effort required to achieve such a result would be huge, so at this time we are just leaving it as a possible future improvement.

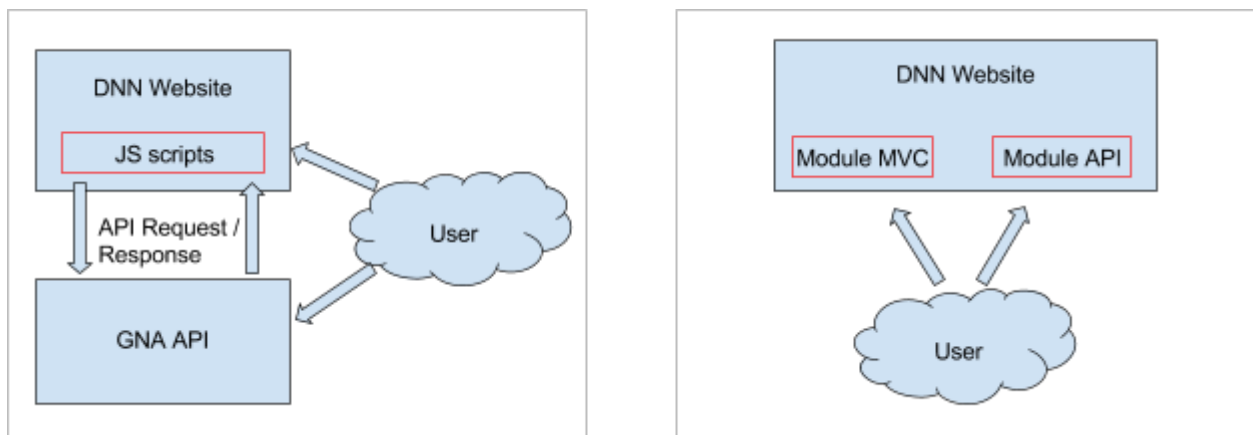
3.3.2 Integration with DNN

DNN is the platform used to publish the INFORM website. While our solution is being developed and it is working as a standalone application, we need to think about how this will work through a DNN website.

As explained before, the GNA app was developed to eventually respond as a remote API also for DNN, which then just needed to have injected and run Javascript code to call the GNA API. This means that the effort for integrating the app is minimal, but it also creates significant issues, because all the methods from the GNA app have to be made available for anyone with no restrictions and this is not what we want.

For these reasons, we propose the new INFORM Tool be fully integrated into the DNN platform. Installing our app as a module on DNN lets us protect all its controllers and methods with DNN built-in **authentication** and **authorisation** features, so that we can choose what users are allowed to do, either when browsing the web pages or when making API calls remotely.

The following picture explains the differences in behaviour between the old GNA app and the new INFORM Tool.





- *Figure 3.1. Split architecture of DNN + GNA system, compared to fully integrated model*

Visible by Administrators only.

User Management

inform

test_inform_new  

E-mail luca.vernaccini@ext.ec.europa.eu

Roles Registered Users,INFORM test,INFORM Partner

System

inform

Workflow Groups

inform2018

Workflows

inform 2018, inform 2018 (no prediction)

- *Figure 3.2: user permission settings*

CreateModel

If you already uploaded country data, you may [send a shape file](#) with spatial data; this let you visualize interactive maps for your Inform release.

Define name for your regional Inform model

Country Data File

No file selected.

Upload a file, or fill the box below with your data.

Note: file formats allowed are XLSX, CSV. If you upload or paste using CSV format, please specify the field separator.

Separator:



Skip first row

- *Figure 3.3. Now users may define their own regional model by uploading data about administrative units and shape files for generating interactive maps*

Create Workflow

New Workflow Name

Model Type Global Regional

Regional Model:

Validity - From

Validity - To

Workflow Group

Or add a new one +

Methodology Template Select a methodology for calculating a composite indicator. You may copy and adapt a methodology of a previous year.

Methodology Description

Comments

Use Prediction Model

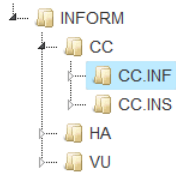
- *Figure 3.4. Definition of a new Workflow — In this example, the model type is REGIONAL*

Methodology Configurator

INFORM 2018 (ID: 405)

It's not possible to save any changes because the methodology is already approved. To modify the methodology, you need to [Un-Approve](#) it first.

Create Rename Delete



Parent	Current Node	Children
CC	CC.INF (Infrastructure)	CC.INF.COM CC.INF.PHY CC.INF.AHC

Details

Name

CC.INF

Fullname

Infrastructure

Select operator type

Aggregation

Select operator

AVG

Default value

-99

VisibilityLevel

2

FamilyGroup

3

CheckPrecision

Indicators for process

Coefficient Indicator	
CC.INF.COM	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
CC.INF.PHY	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
CC.INF.AHC	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>

+ x

CC.INF.COM;CC.INF.PHY;CC.INF.AHC;

Save

- *Figure 3.5. Methodology configurator*

3.4 New Web API

A new web API has been developed to overcome the limits in abstraction and performance that the old one was suffering from.

A new API has been developed for exposing INFORM results, country profiles and all data of public interest.

It will be available on the INFORM website, along with documentation and a query configurator for test purposes. The guideline for understanding how to extract data are as follows: each INFORM model is identified by a 'WorkflowId'. The INFORM models belonging to the same release, are coded with the same 'WorkflowGroupName' (e.g. the INFORM 2017 release and the 5 years of back-calculated models based on the same methodology).

Web API Tester

Compose your search

Inform Model Type

Global

Inform Release

INFORM2017

Workflow

INFORM 2017 v0.3.1

Data Type

Results

Indicators

CC (Lack of Coping Capacity Index)

Geographical Area

Western Europe

Country

Netherlands

Refine search for country

Get Scores

Clear Filters

Active Filters:

Global INFORM2017 INFORM 2017 v0.3.1 Results CC Western Europe Netherlands

Or type custom request URI: /api/

SEND REQUEST

Request URI: /API/InformAPI/countries/Scores/?WorkflowId=261&isoGroup=E2&Iso3=NLD&IndicatorId=CC

```

1 [
  1 {
    i. "Iso3": "NLD",
    ii. "IndicatorId": "CC",
    iii. "IndicatorScore": 1.2,
    iv. "ref_IndicatorId": null,
    v. "ref_IndicatorScore": 0,
    vi. "IndicatorRank": 0,
    vii. "Trend": null,
    viii. "FullName": "Lack of Coping Capacity Index",
    ix. "ShortDescription": "",
    x. "nodelevel": 0,
    xi. "AscDesc": null
  }
]

```

- *Figure 3.6. Web API tester*

3.5 Database migrations

Here follows a summary of the changes applied to the data structure in the latest version of the INFORM web application.

DATA INPUT

```
CREATE TABLE [dbo].[DataInput](
  [Iso3] [varchar](12) NOT NULL,
  [SurveyYear] [int] NULL,
  [PubDate] [datetime] NULL,
  [InsertDate] [datetime] NULL,
  [IndicatorId] [varchar](30) NOT NULL,
  [IndicatorValue] [float] NOT NULL,
  [Note] [text] NULL,
  [Author] [varchar](50) NULL,
  [Source] [varchar](50) NOT NULL,
  [GNAFromDate] [datetime] NULL,
  [GNAToDate] [datetime] NOT NULL,
  [Version] [varchar](100) NULL,
  [IsPredicted] [tinyint] NOT NULL,
  [Timestamp] [datetime] NOT NULL,
  CONSTRAINT [PK_DataInput_New] PRIMARY KEY CLUSTERED
(
  [Iso3] ASC,
  [IndicatorId] ASC,
  [Source] ASC,
  [GNAToDate] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY = OFF,
  ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
) ON [PRIMARY] TEXTIMAGE_ON [PRIMARY]
GO

ALTER TABLE [dbo].[DataInput_New] ADD CONSTRAINT
[DF_DataInput_New_IsPredicted] DEFAULT ((0)) FOR [IsPredicted]
GO

ALTER TABLE [dbo].[DataInput_New] ADD CONSTRAINT
[DF_DataInput_New_Timestamp] DEFAULT (getdate()) FOR [Timestamp]
GO
```

The ObjectId column has been removed and the new primary key is composed of columns for Iso3, IndicatorId, Source, GNAToDate and IsPredicted. IndicatorId and Source column sizes have been slightly reduced in order to have a smaller key size. There are 2 new columns:

- IsPredicted, a 0/1 value that specifies whether the value is predicted or not;
- Timestamp, which is used to show the last update of each row

DATA FINAL

```
CREATE TABLE [dbo].[DataFinal](
  [Iso3] [varchar](12) NOT NULL,
  [SurveyYear] [int] NULL,
  [PubDate] [datetime] NULL,
  [IndicatorId] [varchar](30) NOT NULL,
  [IndicatorScore] [float] NOT NULL,
  [PubType] [varchar](20) NULL,
  [PubDescription] [text] NULL,
  [Note] [text] NULL,
```

```

[Author] [varchar](50) NULL,
[MethodologyId] [int] NOT NULL,
[GNAYear] [int] NULL,
[GNAFromDate] [datetime] NULL,
[GNAToDate] [datetime] NOT NULL,
[Timestamp] [datetime] NOT NULL,
CONSTRAINT [PK_DataFinal_New] PRIMARY KEY CLUSTERED
(
    [Iso3] ASC,
    [IndicatorId] ASC,
    [MethodologyId] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY = OFF,
ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
) ON [PRIMARY] TEXTIMAGE_ON [PRIMARY]
GO

ALTER TABLE [dbo].[DataFinal_New] ADD CONSTRAINT
[DF_DataFinal_New_Timestamp] DEFAULT (getdate()) FOR [Timestamp]
GO

```

As with the DataFinal table, the ObjectId column has been removed and the new key is composed of Iso3, IndicatorId and MethodologyId columns. Version and OidDatainput have also been removed, because they are not relevant and were never used. Finally, the Timestamp column has been added to store the last update of each row.

METHODOLOGY

```

CREATE TABLE [dbo].[Methodology](
    [MethodologyId] [int] IDENTITY(1,1) NOT NULL,
    [WorkflowId] [int] NULL,
    [MethodologyDescription] [text] NULL,
    [MethodologyDate] [datetime] NULL,
    [Author] [varchar](50) NULL,
    [Note] [text] NULL,
    [Status] [varchar](15) NULL,
    [Iso3List] [text] NULL,
    [Version] [varchar](50) NULL,
    [ModelType] [varchar](10) NULL,
    CONSTRAINT [PK_Methodology] PRIMARY KEY CLUSTERED
(
    [MethodologyId] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY = OFF,
ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
) ON [PRIMARY] TEXTIMAGE_ON [PRIMARY]
GO

```

The new column ModelType has been added to specify the model used (global or regional)

WORKFLOW

```

CREATE TABLE [dbo].[WorkFlow](
    [WorkflowId] [int] IDENTITY(1,1) NOT NULL,
    [Name] [varchar](100) NOT NULL,
    [WorkflowDate] [datetime] NULL,

```

```

[FlagMethodologyApproved] [datetime] NULL,
[FlagDataSaved] [datetime] NULL,
[FlagGnaPublished] [datetime] NULL,
[Author] [varchar](50) NULL,
[Comments] [text] NULL,
[GNAYear] [int] NULL,
[System] [varchar](50) NULL,
[WorkflowCompareId] [int] NULL,
[GNAFromDate] [datetime] NULL,
[GNAToDate] [datetime] NULL,
[WorkflowGroupName] [varchar](50) NULL,
[Version] [varchar](50) NULL,
[UsePrediction] [tinyint] NOT NULL,
[Timestamp] [datetime] NOT NULL,
CONSTRAINT [PK_WorkFlow_New] PRIMARY KEY CLUSTERED
(
    [WorkflowId] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY = OFF,
ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
) ON [PRIMARY] TEXTIMAGE_ON [PRIMARY]
GO

ALTER TABLE [dbo].[WorkFlow_New] ADD CONSTRAINT
[DF_WorkFlow_New_UsePrediction] DEFAULT ((0)) FOR [UsePrediction]
GO

ALTER TABLE [dbo].[WorkFlow_New] ADD CONSTRAINT
[DF_WorkFlow_New_Timestamp] DEFAULT (getdate()) FOR [Timestamp]
GO

```

The GNAPeriod column has been removed because it is not relevant anymore. The new column UsePrediction is a self-explained setting, like the Timestamp column already used in previous tables.

COUNTRY

```

CREATE TABLE [dbo].[Country](
    [Iso3] [varchar](12) NOT NULL,
    [IsoGroup] [varchar](12) NOT NULL,
    [CountryName] [varchar](100) NULL,
    [Note] [text] NULL,
    [AdminLevel] [varchar](30) NULL,
    [CategoryType] [varchar](50) NOT NULL,
    [CategoryInfo] [varchar](100) NULL,
    CONSTRAINT [PK_Country] PRIMARY KEY NONCLUSTERED
(
    [Iso3] ASC,
    [IsoGroup] ASC,
    [CategoryType] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY = OFF,
ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
) ON [PRIMARY] TEXTIMAGE_ON [PRIMARY]
GO

```

The Country column has been renamed as CountryName to remove the conflict with the table name.

The new column AdminLevel is meant to store the name of the current administrative level, which could be different for each country / regional model.

Note: all tables with a Timestamp column also have a trigger to automatically update the value

3.6 Index of milestones achieved

3.6.1 Revision of backend code for existing functionalities

- Data Upload
- Data Import from external sources
- Methodology creation and configuration
- Calculation Engine: a major bug relating to data retrieval was found and fixed - and optimisations were made in the code and stored procedures to speed up the process (details to follow in dedicated document).

3.6.2 Update of the database

- Creation of stored procedures + update of existing ones + delete of unused ones
- For migrations, see chapter 3.5

3.6.3 Creation of new regional model (backend + frontend)

- Upload of country data
- Selection of indicators (data input) needed + creation of new indicators (upon approval)
- Upload of indicators data
- Upload of map (shape file) + generation of GeoJson for website

3.6.4 New Web API

This is the public API, developed according to the RESTful paradigm, to expose INFORM results. The name for the new web API will be different to the current one, so third parties will have to change their queries to use the new version. The old one will still be available for a period of time to be defined. See chapter 3.4 for details.

3.6.5 User roles management

Users registered to the website will have different roles and will (or will not) be able to perform specific operations.

3.6.6 Integration with DNN and authentication and authorisation

The application will be installed on DNN as a module and this integration will require some additional tests.

Authentication and authorisation are about the access of the user to the website and to available actions. At this step, a specific level of access will be set for each controller and for each method inside a controller.

GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: <http://europa.eu/contact>

On the phone or by email

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
- at the following standard number: +32 22999696, or
- by electronic mail via: <http://europa.eu/contact>

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website at: <http://europa.eu>

EU publications

You can download or order free and priced EU publications from EU Bookshop at: <http://bookshop.europa.eu>. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see <http://europa.eu/contact>).

JRC Mission

As the science and knowledge service of the European Commission, the Joint Research Centre's mission is to support EU policies with independent evidence throughout the whole policy cycle.



EU Science Hub
ec.europa.eu/jrc



@EU_ScienceHub



EU Science Hub - Joint Research Centre



Joint Research Centre



EU Science Hub

