China Construction Bank Global Green Finance Index Methodology

Introduction

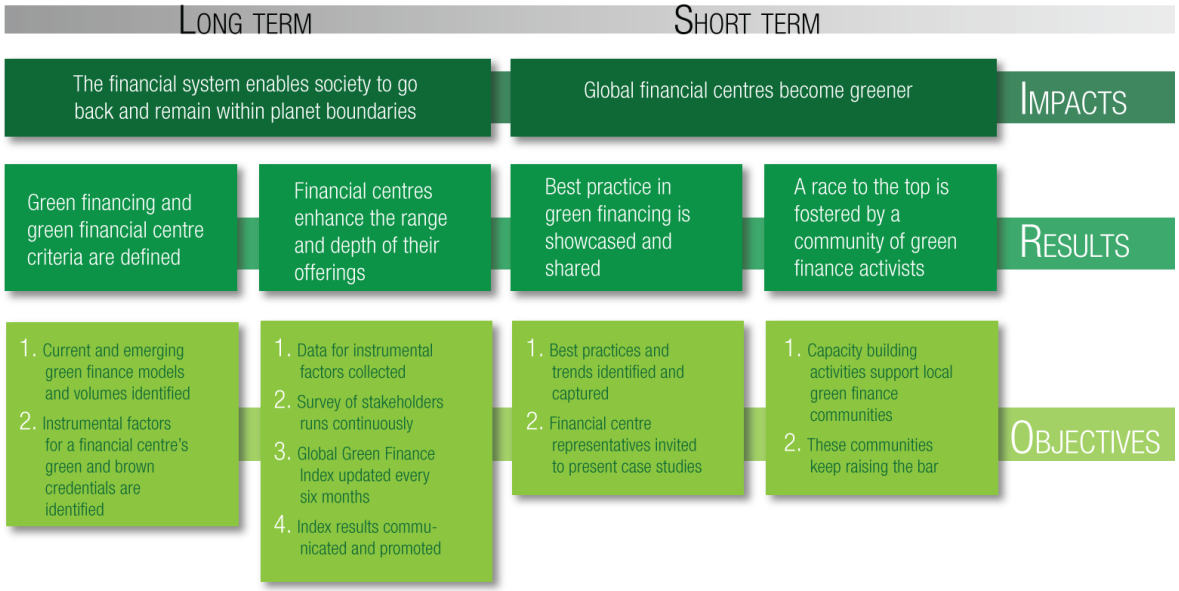
1. This paper sets out the methodology underlying the construction of the China Construction Bank Global Green Finance Index (CCB GGFI). In summary, the process involves taking two sets of ratings – one from survey respondents and one generated by a statistical model – and combining them into a single table. For the first set of ratings, respondents are asked to rate financial centres that they are familiar with on a scale of 1 to 10 in terms of green finance depth and green finance quality. For the second, a machine learning algorithm uses these ratings and a database of other indicators to predict how each respondent would have rated the other financial centres. The respondents' actual ratings and their predicted ratings for the centres they did not rate, are then combined into a single table. The ratings for depth and quality are combined to produce the overall rating for each financial centre.

Background

2. The aims of the CCB GGFI are to encourage financial centres to improve and expand their green finance offering and to focus developments in the financial system in ways that enable society to live within planetary boundaries.

3. The CCB GGFI's aims are set out in the diagram below.

Figure 1: CCB GGFI Aims and Objectives



4. The CCB GGFI is published twice a year and uses qualitative ratings of financial centres' green finance credentials combined with a number of instrumental factors to create an index of financial centres according to their green finance performance.

Approach

Inputs

5. The CCB GGFI provides ratings for the green offering of financial centres calculated by a factor assessment model that uses two distinct sets of input:

- **Financial centre assessments**: using an online questionnaire (http://greenfinanceindex.net/survey), respondents are asked to rate the depth and quality of each financial centre's green finance offering using a ten point scale ranging from little depth/very poor to mainstream/excellent. Responses are sought from a range of individuals drawn from the financial services sector, non-governmental organisations, regulators, universities and trade bodies.
- **Instrumental factors**: these are a range of quantitative data about each financial centre. These 129 instrumental factors draw on data from a range of sources and include:
  - ◊ Sustainability measures, including data on the development of green financial service activities in that centre;
  - ◊ The business environment, including legal and policy factors and statistics on economic performance;
  - ◊ Human capital, reflecting educational development and social factors;
  - ◊ Infrastructure data that reflect the physical attributes of the centre, such as air quality and local carbon emissions, or telecommunications and public transport.

6. A full list of the instrumental factors used in the model can be found here. Due to the way in which the factor assessment model operates, several indices can be used for each area of interest. Neither of these sets of inputs in themselves would allow the creation of a valid index and we use an approach which combines these data in creating the CCB GGFI.

Factors Affecting the Inclusion of Centres in the CCB GGFI

7. The questionnaire lists a total of 130 financial centres which can be rated by respondents. The questionnaire also asks whether there are other financial centres that will improve their green finance offering significantly over the next two to three years. Centres which are not currently within the questionnaire and which receive a number of mentions in response to this question will be added to the questionnaire for future editions.

8. We only give a financial centre a CCB GGFI rating and ranking if it receives a statistically significant minimum number of assessments from individuals based in other geographical locations – at least 25 in CCB GGFI 17. This means that not all 130 centres in the questionnaire will receive a ranking. We will keep this number under review for further editions of the index as the number of assessments increases.

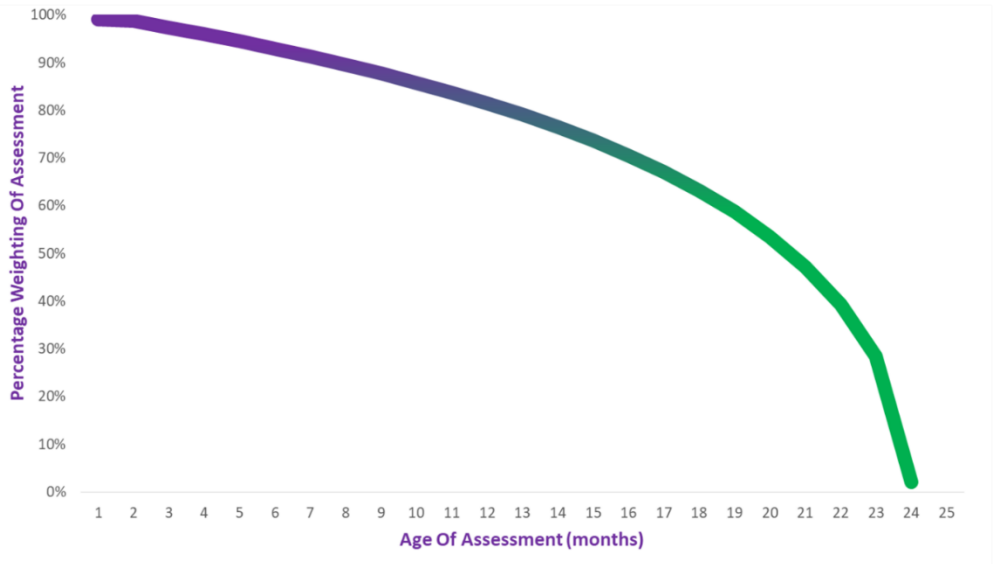
9. We also remove centres from the rankings, if over a 24 month period, a centre has not received a minimum number of assessments.

Financial Centre Assessments

10. Financial centre assessments are collected via an online questionnaire which will run continuously. A link to this questionnaire is emailed to a target list of respondents at regular intervals and other interested parties can complete the questionnaire by following the link given in CCB GGFI publications.

11. For the CCB GGFI:
    - the score given by a respondent to their home centre and respondents who do not specify a home centre are excluded from the model – this is designed to prevent home bias;
    - financial centre assessments will be included in the CCB GGFI model for 24 months after they have been received – we consider that assessments still have validity for a period after they have been given; and
    - financial centre assessments from the month when the CCB GGFI is created will be given full weighting with earlier responses given a reduced weighting on a logarithmic scale as shown in Chart A – this recognises that older ratings, while still valid, are less likely to be up-to-date.

Chart A: Log Scale for Time Weighting



Instrumental Factor Data

12. For the instrumental factors, we have the following data requirements:
    - data series should come from a reputable body and be derived by a sound methodology; and
    - data series should be readily available (ideally in the public domain) and be regularly updated.

13. The rules on the use of instrumental factor data in the model are as follows:
    - updates to the indices are collected and collated every six months;
    - no weightings are applied to indices;
    - indices are entered into the CCB GGFI model as directly as possible, whether this is a rank, a derived score, a value, a distribution around a mean or a distribution around a benchmark;
    - if a factor is at a national level, the score will be used for all centres in that country; nation-based factors will be avoided if financial centre (city)-based factors are available;
    - if an index has multiple values for a city or nation, the most relevant value is used (and the method for judging relevance is noted);
    - if an index is at a regional level, the most relevant allocation of scores to each centre is made (and the method for judging relevance is noted); and
    - if an index does not contain a value for a particular financial centre, a blank is entered against that centre (no average or mean is used).

Factor Assessment Approach

14. Neither the financial centre assessments not the instrumental factors on their own can provide a basis for the construction of the CCB GGFI.

15. The financial centre assessments rate centres on their green finance performance, but each individual completing the questionnaire will:
    - be familiar with only a limited number of centres - probably no more than 10 or 15 centres out of the total number;
    - rate a different group of centres making it difficult to compare data sets; and
    - consider differing aspects of centres' performance in their ratings.

16. The instrumental factors are based on a range of different models and using just these factors would require some system of totaling or averaging scores across instrumental factors. Such an approach would involve a number of difficulties:
    - indices are published in a variety of different forms: an average or base point of 100 with scores above and below this; a simple ranking; actual values, e.g., $ per square foot of occupancy costs; or a composite 'score';

- indices would have to be normalised, e.g., in some indices a high score is positive while in others a low score is positive;
- not all centres are included in all indices; and
- the indices would have to be weighted.

17. Given these issues, the CCB GGFI uses a statistical model to combine the financial centre assessments and instrumental factors.

18. This is done by conducting an analysis to determine whether there is a correlation between the financial centre assessments and the instrumental factors we have collected about financial centres. This involves building a predictive model of the rating of centres' green financial offerings using a support vector machine (SVM).

19. An SVM is a supervised learning model with associated learning algorithms that analyses data used for classification and regression analysis. SVMs are based upon statistical techniques that classify and model complex historic data in order to make predictions on new data. SVMs work well on discrete, categorical data but also handle continuous numerical or time series data.

20. Academic studies have established SVMs as a robust statistical technique, with prediction accuracy rates often well above 90 per cent. Examples of studies on the effectiveness of SVM and explanations of the theory behind the technique can be found at the links below.[1] SVMs have a variety of practical applications in addition to their use in surveys. For example, SVMs can be used to detect faults in diesel engines, or to set the sequence of traffic lights at road junctions. In medicine, they are used to predict whether particular individuals will develop heart disease or diabetes, how well they will recover after a stroke, or what sub-type of cancer they are likely to develop, helping doctors to prescribe more effective treatments.

21. The SVM is run in R, an open source programming language and software environment for statistical computing and graphics that is supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis.

22. The SVM used for the CCB GGFI provides information about the confidence with which each specific rating is made and the likelihood of other possible ratings being made by the same respondent.

---

[1] a. *An Idiot's Guide To Support Vector Machines (SVMs), R Berwick, MIT* (http:// web.mit.edu/6.034/wwwbob/ svm-notes-long-08.pdf)
b. A Gentle Introduction to Support Vector Machines in Biomedicine, Statnikov, Hardin, Guyon and Aliferis (https://www.eecis.udel.edu/~shatkay/Course/papers/UOSVMAlliferisWithoutTears.pdf)
c. *Support Vector Machines, Guenther and Schonlau, Stata Journal* (http://www.schonlau.net/publication/16svm_stata.pdf)
d. Non-linear machine learning econometrics: SupportVector Machine (https://circabc.europa.eu/webdav/CircaBC/ESTAT/ESTP/Library/2017%20ESTP%20PROGRAMME/29.%20Machine%20Learning%20Econometrics%2C%2012%20%E2%80%93%2014%20June%202017%20-%20Organiser_%20DEVSTAT/Materials_Machine_LEUC1502_Module4.pdf)
e. *An Introduction to Statistical Learning, James, Witten, Hastie and Tibshirani* (http://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf)

23. The model then predicts how respondents would have assessed centres with which they are unfamiliar by answering questions such as:

    If a respondent gives Singapore and Sydney certain assessments then, based on the instrumental factors for Singapore, Sydney and Paris, how would that person assess Paris?

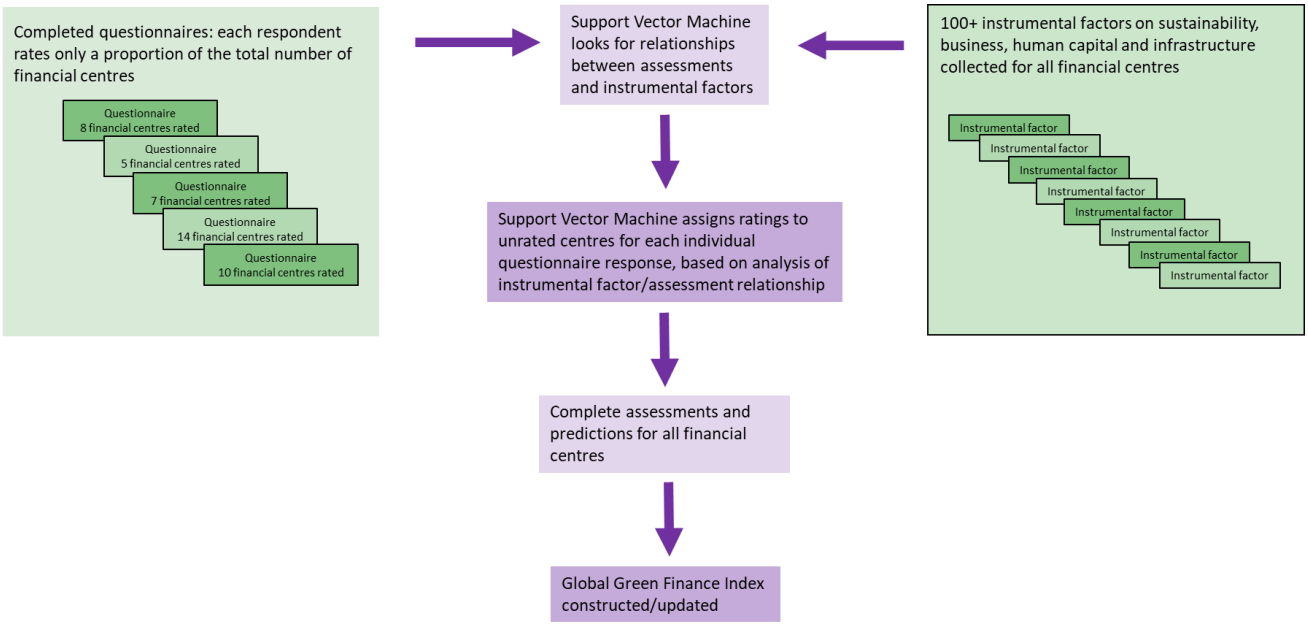
    Or

    If Edinburgh and Munich have been given a certain assessment by this respondent, then, based on the instrumental factors for Edinburgh, Munich and Zurich, how would that person assess Zurich?

24. Financial centre rating predictions from the SVM are re-combined with actual financial centre assessments to produce the CCB GGFI – a set of ratings for financial centres' green finance performance.

25. The process of creating the CCB GGFI is outlined in Chart B below:

Chart B: The CCB GGFI Process



Mathematical Model

26. The mathematical model underlying the process of factor assessment and the creation of the index is described at annex 1.

Validation

27. The rules on data use for both financial centre assessments and instrumental factors are a key part of data validation. In addition, we intend to scrutinise the data set for anomalous patterns, for example, exactly matching assessments given by respondents from the same region or multiple assessments given from the same source. Where there appears to be a discrepancy, we may ignore certain responses in compiling the index.

Updating the Index

28. The CCB GGFI is published twice a year and dynamically updated either by updating and adding to the instrumental factors or through new financial centre assessments.

Use of the Index

29. The CCB GGFI produces a central index rating of financial centres' green finance credentials. The questionnaire collects other data and other analyses of the data are possible, for example:
    - Sector-specific ratings are available using the business sectors represented by questionnaire respondents. This makes it possible to rate different centres in terms of their offering in, for example, insurance as opposed to their strength in debt capital.
    - The factor assessment model can be queried in a 'what if' mode – for example, "how much would Singapore carbon emissions need to fall in order to increase Singapore's ranking against Paris?"

Annex 1: Mathematical Model Used In The Creation Of The Global Green Finance Index

1. This annex describes the model used to create the Global Green Finance Index.

2. The core methodology for creation of the index can be broken down into two parts:

    i. Predicting assessment ratings for financial centres not rated by the survey respondent using support vector machine (SVM); and

    ii. Combining the assessments submitted by survey respondents and the predicted assessments to obtain a comprehensive assessment score for each financial centre under consideration.

Prediction Using Support Vector Machine (SVM)

3. Our training dataset consists of the instrumental factors data (attributes) and assessment scores (the 'label' with discrete integral values ranging from 1-10) for all financial centres being evaluated through the survey.

4. The test dataset consists of instrumental factor data for centres which have not been rated by our survey respondents.

5. The model is 'trained' on the training dataset to predict assessment rating values for all the financial centres in our test dataset.

Implementation

6. We use an R package called 'kernlab' to run 'ksvm' which is one of the many SVM methods available to 'train' the model. It supports classification, regression, native multi-class classification and bound-constraint SVM formulations. Ksvm supports class-probabilities output and confidence intervals for regression.

7. This is an instance of multiclass-classification with k = 10 classes (since our assessment scores can have discrete values ranging from 1-10). Ksvm uses the 'one-against-one' approach in which $\frac{k(k-1)}{2}$ binary classifiers are trained; the appropriate label is found by voting scheme.

Usage

8. The model we use can be described in the following string:

```
Model<-ksvm(CityAssessment~. , data=Train, type= "C-svc", kernel="vanilladot",
C= 0.1, prob.model=TRUE)
```

Parameter description

9. In this string:

- *CityAssessment:* the response vector with one label for each row of x/observation (that is, the assessment score for each financial centre given by the survey respondent);

- *data*: we 'train' our prediction model on the training dataset;

- *type: C-svc (C-classification);* since the labels we are predicting have discrete values we use classification (classification is the problem of identifying to which discrete valued label a test observation belongs to);

- *kernel: vanilladot (linear kernel);* this represents the kernel function used in training and predicting. The training dataset has several attributes (Instrumental Factors) which present a non-linear 'feature space' for prediction. Linear kernel transforms non-linear feature space into linear one for better separability (classification);

- *C: 0.1*; this defines the cost of constraint violation representing the 'C'-constant of regularisation term in Lagrange formulation. The regularization parameter (C) serves as a degree of importance that is given to miss-classifications. The SVM poses a quadratic optimization problem that looks for maximizing the margin between both classes and minimizing the amount of miss-classifications. However, for non-separable problems, to find a solution, the miss-classification constraint must be relaxed, and this is done by setting the mentioned "regularisation";

- *prob.model: true;* set to 'TRUE' builds a model for calculating class probabilities. Fitting is done on output data created by performing a 3-fold cross-validation on the training data and a sigmoid function is fitted on the resulting decision values f.

10. To obtain the class probabilities for our multi-class classification model just computed, we set the parameter type to 'probabilities' in the 'predict' function in R to

    Pred<-predict(Model, Test, type="probabilities")

    This gives probability value for each assessment score value (1-10) an individual might give a centre. The expected value ($\sum p_i x_i$) calculated for each row in this dataset gives us the assessment rating values for the observations (financial centres) in the test dataset.

Index Creation

11. The second step in the creation of the index consists of:
    - calculation of interim weighted scores using probabilities obtained through the SVM process; and
    - combining the assessment scores used in our training data with predicted assessments.

Interim weighted score calculation

12. As further editions of the index are produced, we intend to predict assessment ratings for all financial centres under consideration and only consider questionnaire assessments submitted in past two years. To account for the age of the data, we will multiply the assessment ratings calculated for financial centres in the test dataset by log value of time (calculated from the date the survey was submitted by respondent) to calculate a weighted assessment rating for each centre. The more recent assessments receive a higher weight than older ones.

13. Where a new financial centre is added to the questionnaire, predictions created for that centre based on questionnaire responses given before its inclusion will be multiplied by 0.85 to reduce the weighting of that assessment. Responses made after the date of the centre's inclusion will be given full weighting.

14. Similarly, the assessment scores in our training dataset (assessments for centres given by the survey respondents) are multiplied by discounted log value of time.

Combining the assessments used in our training data with predicted assessments

15. The final step is calculating a weighted mean score for each financial centre by summing all the weighted ratings (from training and test dataset) calculated for that financial centre and dividing by the sum of log discount. This score value is multiplied by 100 to

get the financial centre rating. According to the ratings we index the financial centres and generate individual rankings.

References:

1. https://cran.r-project.org/web/packages/kernlab/kernlab.pdf
2. https://escience.rpi.edu/data/DA/svmbasic_notes.pdf