Using Markov chains to predict trends in Rwanda's school system

PROOF OF CONCEPT

MARCH 2020

laterite
DATA | RESEARCH | ADVISORY

Summary

The structure of Rwanda's basic education system is asymmetric, with many more students enrolled in the early grades than in the later grades. According to 2018 data from the Rwandan ministry of education (MINEDUC), student enrollment in primary 1 (517,243) was about double the enrollment in primary 6 (260,060) and eleven times the number of children at the send of secondary 6 (44,437). This lopsided structure is the result of a rapid increase in the intake of students at primary 1 level and the balance of progression, dropout and repetition observed in each cohort of students. All these factors combined mean that predicting the number of students attending a grade next year is more complex than just moving a cohort along.

This proof of concept demonstrates that it is possible to use Markov chains and available data to create evidence-based models to support planning and decision making. Our approach deploys Markov chains to model the enrollment structure in the Rwandan school system using data from MINEDUC statistical yearbooks and Laterite's report on Dropout and Repetition in Rwandan schools (Laterite 2017).

Markov chains are statistical models that describe a sequence of future events based on a set of probabilities defined by previous events. For example, two students enrolled in Secondary 1 in 2018 will have different chances of progressing, dropping out or repeating according to their gender, age, socio-economic background and past academic performance. This means that in 2019 the students may be again classmates enrolled in Secondary 2, or maybe one of them stayed in Secondary 1. These new parameters will again influence their new transition probabilities in 2020. By scaling up the Markov chain from the individuals to the school population, it's possible to estimate the number of students enrolled per grade per year in Rwanda.

Our proof of concept forecasts the structure of the Rwandan school system up to 2024. The results suggest that the secondary school population will almost double in the next four years, while enrollment rates in primary school will remain at today's values. This has implications for resource allocation & planning across schools in Rwanda, specifically on the number of teachers that will need to be trained if current pupil-to-teacher ratios are to be maintained.

Proof of concept

Context

The expansion of Rwanda's education system has come hand in hand with a swelling of the student population in the early grades. Rwanda's almost universal access to schooling means that primary 1 is replenished every year with new students. On the other hand, students have not been progressing through the education system as expected due to high repetition rates in the early grades. These two factors combined lead to a disconnect between age and grade: many students are older than expected for the class they share with smaller children.

The lopsided structure of the education system and the heterogeneity of class composition (in terms of age and other factors) make it difficult to estimate future student population levels by grade. Students from different age groups have very different probabilities of promotion, repetition, dropout, and re-entry. In addition to age, regional variations between rural and urban areas, socio-economic status and gender also play a role in the individual trajectory of a student. All these factors influence how children progress through the education system. For example, a child that has repeated a grade is much less likely to repeat that grade again; and children that have repeated multiple times are more likely to drop out.

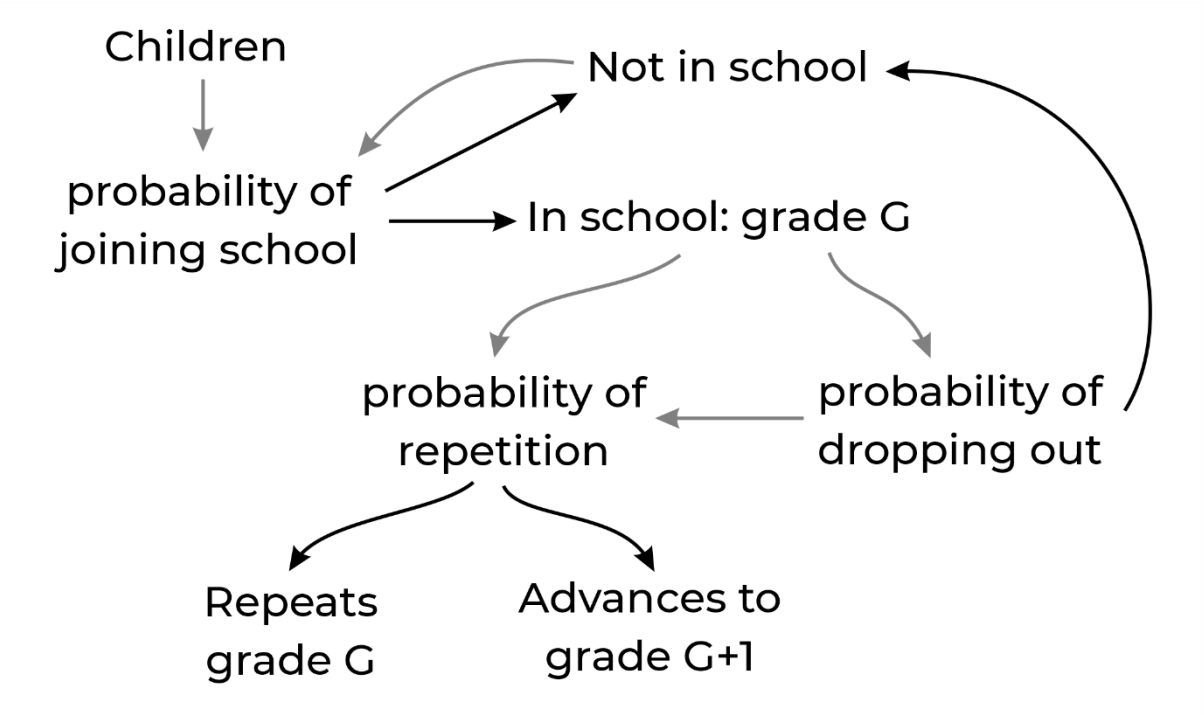
Estimating the future student population has important implications for resource allocation. For example, it informs the number of teachers Rwanda needs to start training today to meet its teacher-to-pupil ratio targets in the future.

Modelling Rwanda's education system with Markov chains

We propose using Markov chains to understand how the current structure of Rwanda's education system influences its future trajectory. In the absence of large policy changes or shocks to the education system, Markov chains can predict how a system will look based on its current state and on a set of transition probabilities. This approach has been used before to help forecast what education systems will look like in the future (see e.g., Abimbola, 2014; Egbo et al., 2018). Transition probabilities in this case refer to the rates at which students might get promoted, repeat, dropout or re-enter the education system. These probabilities vary depending on characteristics that can be assumed to be stable over time, such as gender, location, and the wealth quintile of their families; and on characteristics that vary over time as the student progresses through the education system, such as the grade the student is enrolled in, her age, and her previous repetition history.

In this proof of concept, we focus on predicting future enrollment levels by grade.

Figure 1 / The probabilities that influence a student's transition through the education system



Implementing Markov chains to model trends in enrollment

The core components of a Markov chain are:

(i) *a starting point*, in this case the state of the education system at a given point in time; and

(ii) *the transition probabilities*, the student probabilities to get promoted, repeat, dropout or enter/re-enter the education system.

This proof of concept uses two separate Markov chain models, providing us with different starting states and transition probabilities: (i) one based on the EICV4 data, with 2012 as the base year; (ii) the other using EICV5 data with 2015 as the base year. The data for this analysis was extracted from the Integrated Household Living Conditions Survey (or Enquête Intégrale sur les Conditions de Vie des ménages – EICV) datasets of 2013/14 (EICV 4) and 2016/17 (EICV 5). These datasets contain a detailed set of survey questions on education and capture the characteristics of young Rwandans who are either enrolled or not enrolled in school. Working with EICV4 allows us to test predictions against the reality of the education system between 2014 and 2018. Using EICV5 allows us to account for the most recent changes in the structure of the population and in the transition rates. Combining the two models mitigates over-fitting.

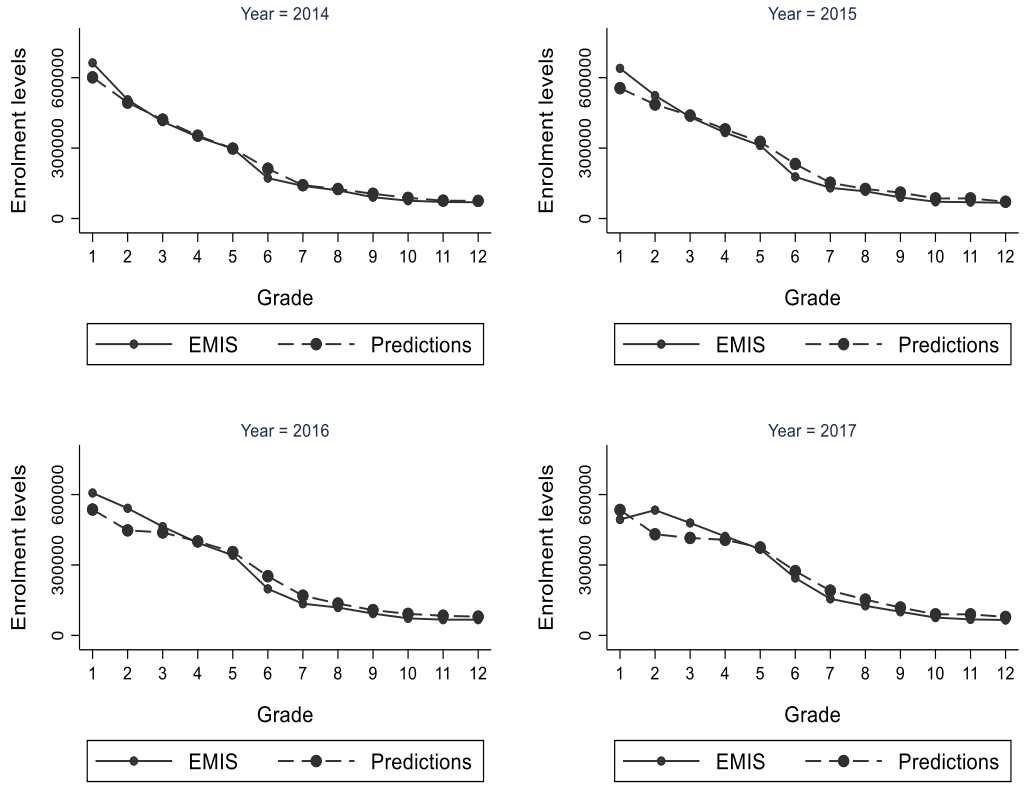
The Markov chain takes the EICV data as a starting point and then updates this data for each predicted year according to the student's transition probabilities, which depend on: (i) their grade, (ii) gender, (iii) location (rural/urban) and (iv) the wealth quintile of their families. For example, a student aged 16 and enrolled in Primary 6 (P6) has a 65% chance of progressing, a 30% chance of dropping out and a 5% chance of repeating. This means that we expect the student to get promoted in 65 out of 100 runs of the model, to drop out in 30 out of 100 runs and to repeat in the 5 remaining runs. In the subsequent year, this student will now be aged 17 and either enrolled in P6, Secondary 1 or out-of-school. These new parameters will influence the new transition probabilities for this student.

Model tuning and robustness

One way to check whether the Markov chain models yield realistic results is to compare predictions against existing data from Rwanda's education system. Our models use individual data from national surveys as input, the first with a base year in 2012, the second with a base year in 2015. We can compare predictions from our models to official data from MINEDUC (called EMIS) collected and published each year as a statistical yearbook. The latest publicly available yearbook is from 2018. Figure 2 compares predictions from our Markov chain model with no adjustments to the administrative data from MINEDUC's EMIS system.

We find that differences between the Markov chain models and MINEDUC data relate to: (i) discrepancies between survey and administrative data; and (ii) policy changes that have affected transition rates. Empirical differences between administrative and survey data are not an uncommon occurrence (Penneck, 2007; Johnson and Moore, 2008; Figlio et al., 2016; Adriaans et al., 2019). For example, administrative data may overestimate entry into the education system when students are officially enrolled but then not attend; survey data can overestimate promotion rates when children shy away from saying that they have dropped out. Policy changes that are not captured in the models, but that have affected transition rates include: a large push to increase enrollment during the 2014-2016 period, during which primary school enrollment increased from 2.39 to 2.54 million children; and a drop in repetition rates in 2016-2017, after the enforcement of MINEDUC's policy to limit repetition rates as much as possible.

Figure 2 / Comparing the Markov chain model (based on EICV 4) with no adjustments to MINEDUC data (EMIS)



We account for these differences with a set of adjustments to the models: (i) we uniformly increase the probability of entry for children that have not yet started their education; (ii) we uniformly increase the probability of dropout for children above the age of 12; and (iii) we adjust the probability of repetition in the transition from 2016 to 2017. These changes increase the risk of over-fitting, but on the other hand these changes are simple multiplications by a known factor and they do not increase the complexity of the model.

To strengthen predictions and reduce the risk of over-fitting we:

(i) **Simulate the model multiple times** by introducing random errors in the predictions to get a distribution of enrollment outcomes. This allows us to estimate the statistical confidence interval around our predictions.

(ii) **Base our estimates off two difference Markov chain models**, with different starting points and transition probabilities.

Results

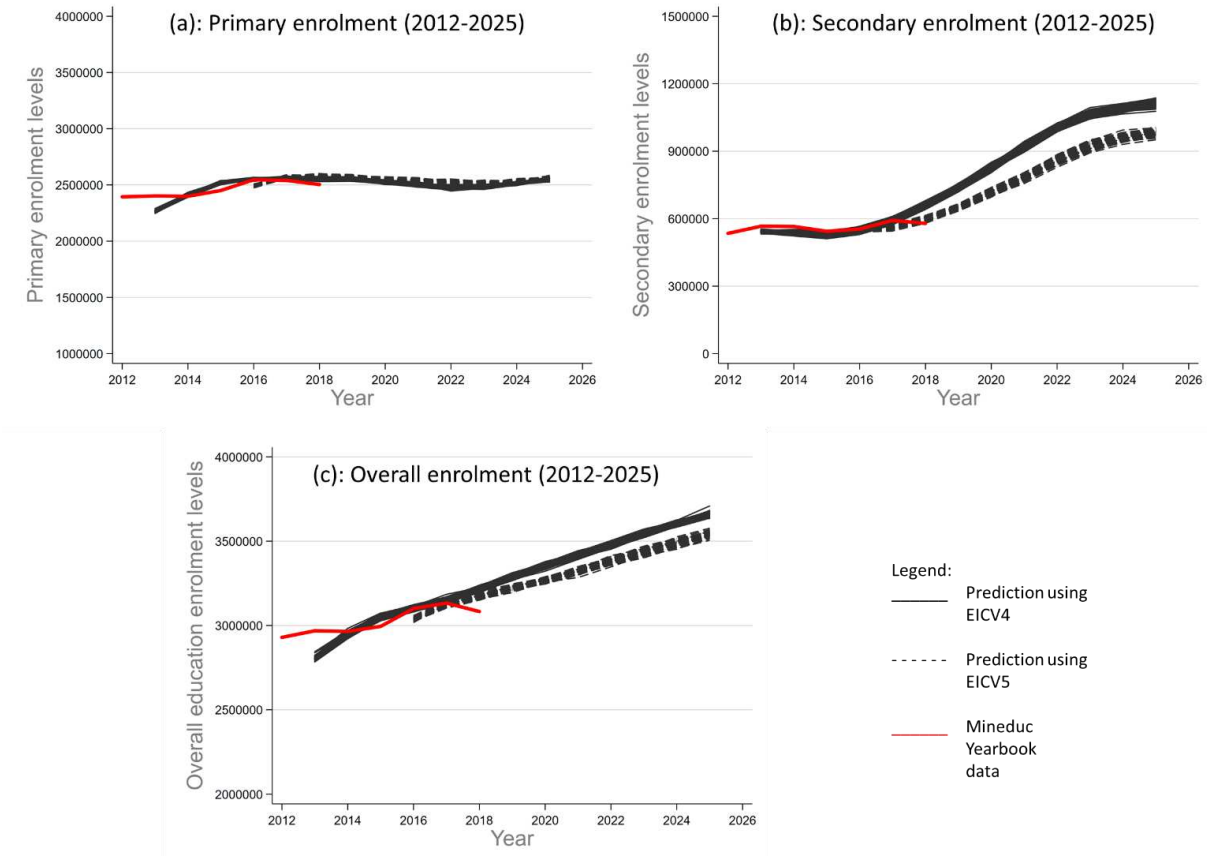
The trends predicted by our Markov chain proof of concept point towards stable primary school enrollment levels. According to our model, enrollment levels in primary education are poised to stay relatively stable as a whole, hovering around the 2.5 million

students mark between 2018 and 2024. The model anticipates that there will be a greater equilibrium in the distribution of students by grade and that more students will be making it through to Primary 6.

On the other hand, secondary school enrollment levels are predicted to increase from about 660,000 students in 2018 to over 1 million students in 2024 (see Figure 3 and Table 1). The model predicts that the grades that will face the greatest resource constraints in the very future are Secondary 3 (S3) to Secondary 6 (S6) where enrollment levels increase dramatically by 2024. This is particularly severe for S3, where the model predicts 195,000 students in 2024, compared to the 100,000 registered in 2018.

This projected increase in the number of secondary students has very important resource implications. The current pupil-to-qualified teacher ratio for secondary is 28:1. To maintain this target over time the secondary school system will need 16,400 more teachers between 2018 and 2024. This figure assumes that all current secondary teachers (22,966 in 2018) stay in service and does not take into account teacher turnover or the need to replace teachers exiting the education system.

Figure 3 / Predictions of enrollment levels using EICV4 and EICV5 compared to the MINEDUC Statistical Yearbook



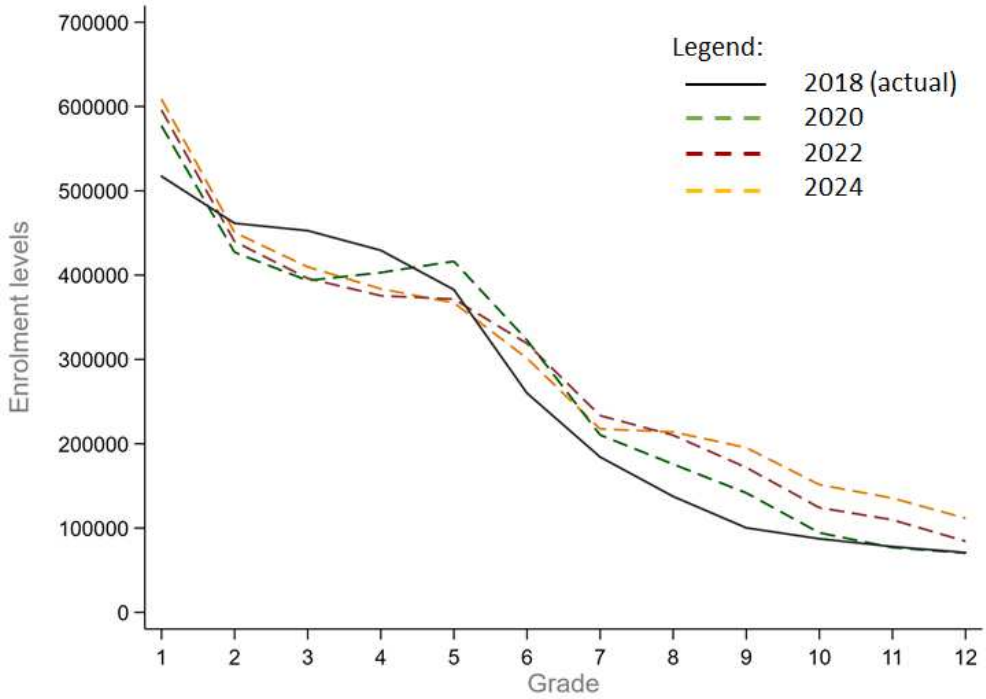
Note: The thickness of the predicted plots represents the variation in the predictions over 100 simulations.

Table 1 / Actual enrollment and enrollment figures predicted by the proof of concept with 5th and 95th percentile distribution of the estimates in brackets

Grade	Actual enrollment in 2018	Predicted enrollment in 2020	Predicted enrollment in 2022	Predicted enrollment in 2024
Primary 1	517243	577181 (568021, 587147)	595776 (585384, 604595)	608682 (596705, 619460)
Primary 2	461499	427040 (417624, 436491)	439808 (430943, 450223)	450545 (440950, 458663)
Primary 3	452745	393593 (385197, 401653)	396158 (385859, 405699)	409713 (399239, 419389)
Primary 4	429412	403037 (393311, 413360)	375552 (367605, 384717)	383730 (376070, 393339)
Primary 5	382746	416384 (406746, 425105)	371570 (364272, 380349)	367146 (356671, 375599)
Primary 6	260060	322953 (313794, 331779)	318920 (311499, 326753)	301385 (292657, 310723)
Total Primary	**2503705**	**2540189**	**2497785**	**2521201**
Secondary 1	184327	210470 (204534, 218828)	233456 (225648, 240859)	217510 (212074, 225442)
Secondary 2	137503	175756 (169315, 181660)	210087 (202317, 218109)	213998 (207165, 221661)
Secondary 3	100263	141644 (135690, 146594)	171780 (165722, 177533)	195259 (188762, 202495)
Secondary 4	87305	94436 (90519, 98412)	124144 (118440, 128966)	151444 (144301, 157067)
Secondary 5	77987	76938 (73114, 81818)	109756 (105072, 114422)	135489 (129873, 141052)
Secondary 6	70900	70271 (66675, 74120)	84308 (80200, 88844)	111615 (106413, 116847))
Total Secondary	**658285**	**769516**	**933531**	**1025314**
Total Primary and Secondary	**3161990**	**3309705**	**3431316**	**3546515**

Source of actual enrollment data: MINEDUC/EMIS data

Figure 4 / 2020-2024 predictions per grade versus observed student count in 2018. Primary school comprises grades 1 to 6; secondary school grades 7 and above



Limitations and directions for future research

This proof of concept demonstrates how Markov chains can be used to model enrollment structure and support evidence-based planning in the education sector. Like any model, Markov chains come with limitations that are important to keep in mind, including:

- **Markov chain models assume that transition rates will remain stable over time.** These models do not consider future and external interventions that can have an impact on the whole system, such as government policies to reduce repetition rates.
- **Here, transition probabilities are calculated based on information from one time period only (for both the EICV 4 and EICV 5 models).** We would have obtained more robust transition estimates with a greater number of time periods to train on.
- **The structure of the data does not allow us to test the quality of predictions.**
- **The fine-tuning of transition probabilities was not data driven.** The tuning of the models we presented was conducted empirically, in such a way that predictions produced a good fit with actual MINEDUC data. This does increase the risk of over-fitting the model.

Our future research in this area will focus on: (i) strengthening model design and the testing thereof; and (ii) using Markov chain models to deliver new insights to policy makers in the education sector.

METHODOLOGICAL NOTE

Transition probabilities for the Markov chain models were estimated using linear probability models. The dependent variables for the models comprised a pupil's age, gender, whether they have repeated previously, wealth quintile, grade, and location. After estimating the probabilities of dropout, repetition, re-entry, and enrollment, the model transforms these probabilities into binary outcomes by generating random variables and deciding outcomes based on whether the random variable was less than or more than the probability of a certain outcome. For example, if the predicted probability of a pupil to drop out is 0.67 and the random variable's value is 0.3 then the model determines that the pupil will drop out.

Further, the linear probability models that were used to calculate the different outcomes also allow us to calculate prediction errors for each outcome. We use these errors to obtain a distribution of predictions for each pupil over multiple runs of the model.

To get the confidence intervals of enrollment levels, we use the variation introduced into the process by the prediction error of the linear probability model and the random variable that is used to binarize outcomes. This variation allows us to simulate the model multiple times and attain a distribution of various possible enrollment levels over the years.

REFERENCES

Abimbola, K. (2014). *Application of Markov chain model to Educational System: The use of Markov Chain in Educational System*. LAP LAMBERT Academic Publishing.

Adriaans, J., Valet, P., Liebig, S. (2019). Comparing administrative and survey data: Is information on education from administrative records of the German Institute for Employment Research consistent with survey self-reports?. *Quality & Quantity*. https://doi.org/10.1007/s11135-019-00931-4.

Egbo, M. N., Bartholomew, D. C., Okeke, J. U., & Okeke, E. N. (2018). Markov Chain Approach to Projection of Secondary School Enrollment and Projection of Teachers. *Open Journal of Statistics*, 8(03), 533.

Figlio, D., Karbownik, K., Salvanes, K.G. (2016). *Chapter 2 - Education Research and Administrative Data*, Editor(s): Eric A. Hanushek, Stephen Machin, Ludger Woessmann, Handbook of the Economics of Education, Elsevier, Volume 5, 2016, Pages 75-138, ISSN 1574-0692, ISBN 9780444634597, https://doi.org/10.1016/B978-0-444-63459-7.00002-6.

Johnson, B., & Moore, K. (2008). Comparing Administrative and Survey Data. *IRS Statistics of Income Working Paper Series.*

Laterite (2017). *Understanding Dropout and Repetition in Rwanda*. Kigali, Rwanda. url: http://www.rencp.org/wp-content/uploads/2018/09/DROPOUT-STUDY-FULL-REPORT.pdf.

Penneck, S. (2007). Using administrative data for statistical purposes. *Econ Lab Market Rev* 1(10). https://doi.org/10.1057/palgrave.elmr.1410152.

laterite

DATA | RESEARCH | ADVISORY

From data to policy

Rwanda | Ethiopia | Kenya | Uganda | The Netherlands

www.laterite.com