MEMORANDUM

505 14th Street, Suite 800
Oakland, CA 94612
Phone: 510-285-4647
Fax: 510-830-3701
cksoll@mathematica-mpr.com
www.mathematica-mpr.com

TO: Julian Glucroft

FROM: Chris Ksoll, Seth Morgan, and Randall Blair **DATE:** 8/29/2019

SUBJECT: Recommendations for ADP public and restricted access files

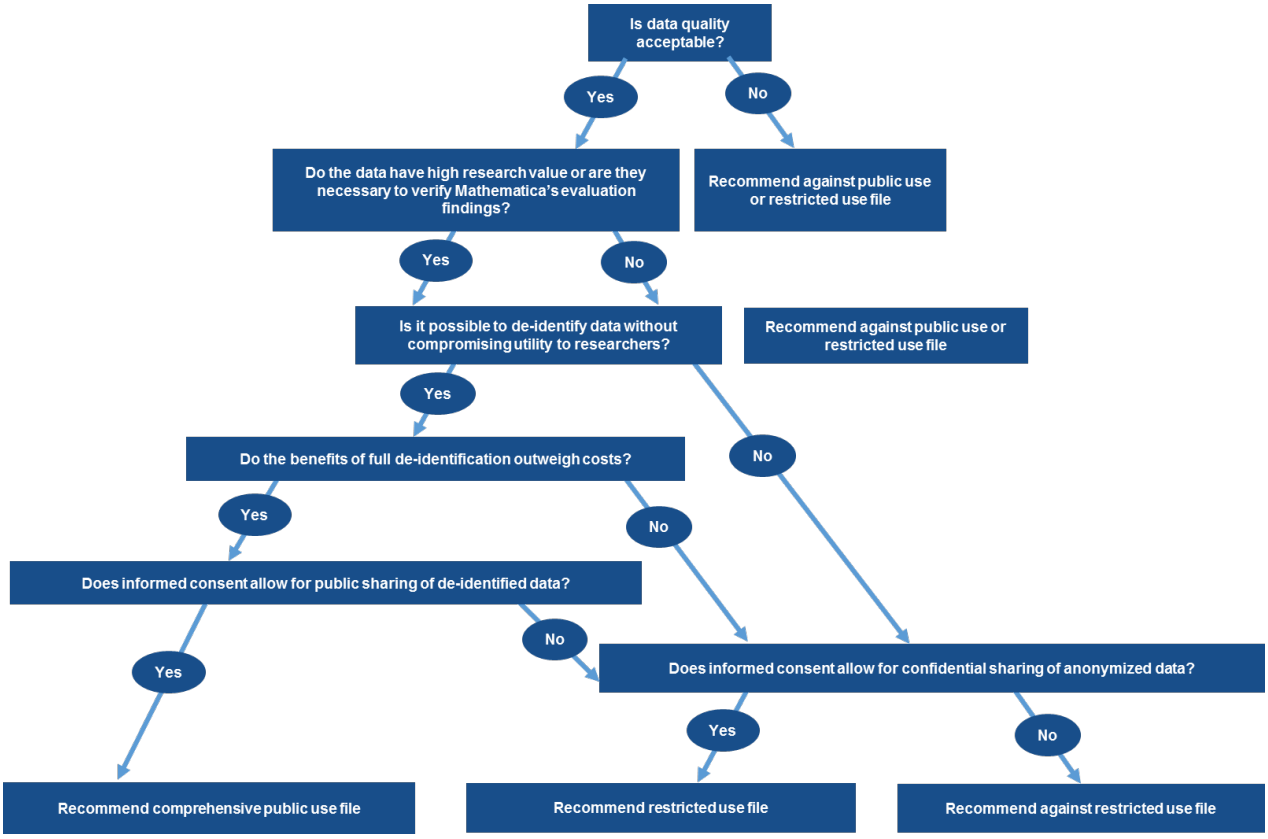
Following discussions with MCC, this memo outlines our proposal for creating public use and restricted access files for ten quantitative data sources related to evaluations of the Agriculture Development Project (ADP) in Burkina Faso. We base these recommendations on several factors, including the overall quality of the data, the potential use of data sources to verify Mathematica's impact findings, and informed consent language. In Figure 1, we outline the decision tree we used in making these recommendations.

Recommendations

Applying the decision tree from Figure 1, we first assessed whether each data source was a good candidate for a comprehensive public use file with fully de-identified[1] data. Good candidates would meet all five of the following conditions: (1) data quality is acceptable, (2) data have high research value, (3) full de-identification for public use files would not compromise utility to researchers, (4) the benefits of a public use file *over and above* a restricted access file—namely the increased accessibility of files to outside researchers—justify the cost of de-identification, and (5) informed consent language permits sharing of de-identified data. *We do not recommend submitting a comprehensive public use file for any of the ten ADP data sources because none of these sources meet all five criteria.* Table 1 provides a summary, assessment and recommendations for each data source.

[1] To de-identify is to remove all direct identifiers from a dataset *and* to drop or mask additional variables—referred to as indirect identifiers—that either alone or in combination with other variables would allow someone to re-identify a respondent.

Figure 1. Decision tree used to make recommendations on ADP data



However, we recommend submitting a restricted use file for five of the ten data sources. These sources meet the criteria above concerning data quality, research value, and consent language, but full de-identification is either infeasible or too costly to justify the construction of a comprehensive public use file. These data sources are the (1) Di PAP baseline survey, (2) Di Lottery baseline survey, (3) farmer training baseline household survey, (4) livestock ("barymetric") survey, and (5) farmer training supplemental household survey (IMPAQ). For these data sources, we propose to submit complete datasets with direct identifiers removed, but not fully de-identified, to the Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan, as permitted by the consent language. ICPSR has a submission option that allows for data-sharing "under a set of highly controlled conditions to approved researchers" with a data protection plan, detailed research plan and IRB approval. These safeguards for accessing data are appropriate given the high risk of re-identification in the

ADP datasets. Based on ICPSR guidelines, submission to ICPSR would require MCC's explicit permission.[2]

For three of these five data sources, we propose to submit key indicator public use files as a complement to restricted access files, as such public use files are permitted by consent language and offer a low-cost alternative to fully de-identified public use files. Key indicator public use files would contain a limited set of variables that are either (a) critical to verifying evaluation findings, such as key outcome indicators, or (b) of particularly high value to outside researchers.[3] The files would include key attributes (such as access to irrigation or type of plot used) without which the data is not useful for research, but would exclude geo-referencing data and other characteristics that require extensive de-identification. We list the proposed indicators for the key indicator public use files in Tables 3-5. Each key indicator public use file would also include the full list of variables in the restricted access file to highlight what additional data is available if researchers meet the requirements for restricted access. As such, key indicator public use files would help outside researchers determine whether they should apply for restricted access to complete data sets.

We recommend that five of the ten data sources are not submitted as restricted access files or public use files, given their low quality or limited utility for third-party researchers. These five data sources are the (1) farmer training baseline crop yield survey, (2) farmer training baseline crop yield survey for monitoring purposes, (3) farmer training fishing survey, (4) farmer training institutional survey, and (5) farmer training interim crop yield survey.

Next steps

We propose that MCC review our recommendations and make a final determination on each recommendation. We also request that MCC provide us with written permission to submit data to ICPSR, and that MCC validate the indicator sets for each of the three proposed key indicator public use files. The final decisions for the Di PAP and Di Lottery surveys could be delayed until the final round of data collection—currently scheduled for 2020—because we will have more complete information on all rounds of the data at that point.

---

[2] We also note that the consent of the Di Lottery baseline and Di PAP datasets promise respondents that the data collected will only be used for research purposes. This language requires restricting access to the data along the ICPSR criteria for restricted access.

[3] Examples of high-value variables include livestock weights from the barymetric survey, as such variables are uncommon in datasets from developing countries.

Table 1. Data files used in ADP evaluations

Activity	Evaluation and survey	Data collector and year	Objective and description of survey	Sampling and sample size	Modules / Content	Assessment	Recommendation
WMI	Di PAP compensation land lost and compensation information; Di PAP baseline survey.	BERD,2011, 2012, 2013	Compensation data on all PAPs collected for compensation purposes. Retrospective baseline survey collected for a sample of PAPs as some PAPs began receiving their new plots on the perimeter. BERD developed the baseline survey to collect baseline data for a pre/post analysis. Unfortunately, the baseline data did not capture information on the value of agricultural output prior to the relocation, thus preventing a pre/post analysis. The data was collected between October 3, 2013 and October 19, 2013.	N=500, selected by BERD as a representative sample of PAPs (roughly 1500 total). Sampling information is incomplete. A total of 388 PAPs out of the selected 500 PAPs completed the survey resulting in a non-response rate of 22.4 percent.	Household demographics Production and land use outside perimeter. Land use in perimeter, revenue, household assets, Perspective on compensation Access to credit, Revenue, Training received	Data **quality is acceptable.** Data have **moderate research value**, particularly to verify evaluation findings. **Data de-identification is feasible, but would be costly** for the purposes of comprehensive public use file, given high identification risk. Consent language in Mathematica's interim survey allows for administrative records from the interim sample to be **shared confidentially with direct identifiers removed, as well as de-identified and shared publicly**. Since consent language is not available for Di PAP compensation data, we cannot share these data for non-interim sample respondents. Consent language for the baseline survey is not available, but the survey states that all data will be "strictly confidential and only used for research purposes". IRB guidance permits submission of this dataset to ICPSR.	Submit data for interim sample as **restricted access file** to ICPSR. Also submit Di PAP baseline dataset as **restricted access file** given updated IRB guidance. Submit **key indicator public use file** for interim sample for key interim and final Di perimeter outcome indicators that allow for ERR calculation – this primarily includes profits (Table 3). *Note: Because the baseline data has only very limited information and unclear sampling, the key indicator public access file will be constituted from Mathematica's interim and final surveys; the baseline data mainly contribute context.*

Activity	Evaluation and survey	Data collector and year	Objective and description of survey	Sampling and sample size	Modules / Content	Assessment	Recommendation
	Di Lottery applicant data and baseline survey	BERD 2013; CERFODES, 2013	Di Lottery applicant data was collected for beneficiary selection purposes, and was publicly posted for verification. Baseline survey collected information from 2,178 applicants who met basic eligibility criteria at the start of the baseline survey (verification of eligibility data was delayed). This includes many of the same variables as applicant data as well as data on basic demographics, socioeconomic status, and other background characteristics that are used to verify the comparability of the two groups.	The 2,178 applicants who met the eligibility criteria for the lottery (before the end of the restitution period) were surveyed (with high response rates).	Household demographics and assets; Experience in agricultural production of candidate and household; Land cultivated by applicant and household; Household revenue	Data **quality is acceptable.** Data have **high research value**, particularly to verify impact findings. **De-identification is not feasible** given high risk of identification (Applicant information was publicly posted). Consent language allows data to be **shared confidentially with direct identifiers removed**.	Submit **restricted access file** to ICPSR.
DA	Farmer Training baseline household survey	NORC and CERFODES, 2011 (covering 2010-2011 dry and 2011 rainy seasons)	Provide a baseline to evaluate a range of ADP activities using a matched comparison group design. Therefore data was collection in 30 intervention and 60 comparison villages. The first round of the baseline data was collected immediately after the dry season of 2011. The second round was collected immediately after the rainy season of 2011.	Households were listed in the 30 villages of the two ADP intervention provinces (Sourou and Comoé) and 60 villages of the 17 provinces selected for comparison. The listing information was used to match each intervention village household to a similar household in one of the comparison villages, and 1,082 matched pairs (a total of 2,164 households) were randomly selected.	Household Agriculture Animal husbandry Forestry Consumption and credit Food security Health	Data **quality is acceptable** Excluding animal husbandry, forestry, food security and health modules, data have **high research value** to verify impact findings. **Moderate de-identification required**, given identification risks. Data would be **costly to fully de-identify**, given the survey length and easy identification of Comoé basin respondents. Consent language would allow data to be **shared confidentially with direct identifiers removed, as well as de-identified and shared publicly.**	Submit **restricted access file** to ICPSR Submit **key indicator public use file** that includes yields for focus crops, total agricultural income and use of agricultural technologies promoted by MCC. This will allow for pre-post analysis (Table 4). *Note: For the restricted access file, we propose to conduct limited data preparation for animal husbandry, forestry, food security and health modules, as they do not relate to the evaluation and no panel data will be available.*

Activity	Evaluation and survey	Data collector and year	Objective and description of survey	Sampling and sample size	Modules / Content	Assessment	Recommendation
	Farmer training crop yield survey	NORC and CERFODES, 2011 rainy season	Crop yield measurements were collected on a subsample of the household survey sample as the self-reported crop yields from the household survey were deemed unreliable.	170 pairs of treatment and comparison households (340 households total) were randomly selected from the household baseline survey sample	single module- yield squares	Data **quality is low** based on the "quality assurance report by Direction Générale de la Promotion de l'Economie Rurale (DGPER), the national office for agricultural statistics in Burkina Faso that examined the data. DGPER concludes that the rice and maize crop yields data are of inacceptable quality and should not be used." (Unpublished data quality report, IMPAQ 2014) We do not have questionnaires or documentation on the methodology used in implementing the crop yield survey or data cleaning.	We propose neither a restricted access file nor a public use file, given low data quality.
DA	Farmer training crop yield survey (for program monitoring purposes)	NORC and CERFODES, 2010-2011 dry season (data collected May-June 2011) 2011 rainy season (data collected Dec 2011-Jan 2012)	NORC/CERFODES used the same yield square methodology to implement a second crop yield survey on a separate sample of farmers in treatment villages for program monitoring purposes.	Respondents were selected from farmers in 65 production sites—defined as a concentration of farmers, such as a group of farms which surround a dam, riverbank, borehole, or well—who reside in the intervention villages. Sampling frame: 3,308 farmers listed during the dry season, 3,725 during the rainy season. Sample: 85 farmers with 167 crop yield measurements for the dry season; 143 farmers with 159 crop yield measurements for the rainy season.	single module- yield squares	Data **quality is low** based on the "quality assurance report by Direction Générale de la Promotion de l'Economie Rurale (DGPER), the national office for agricultural statistics in Burkina Faso that examined the data. DGPER concludes that the rice and maize crop yields data are of inacceptable quality and should not be used." (Unpublished data quality report, IMPAQ 2014). (See assessment for Farmer training crop yield survey)	We propose neither a restricted access file nor a public use file, given low data quality.

Activity	Evaluation and survey	Data collector and year	Objective and description of survey	Sampling and sample size	Modules / Content	Assessment	Recommendation
	Farmer training fishing survey	NORC and CERFODES, 2011-2012 (October 2011-February 2012)	This survey was collected for monitoring and not evaluation purposes. The data were only collected in the treatment areas, for a brief period, on a sample independent of the household baseline survey sample.	Five fishermen from 35 fishing sites—defined as places where fishing is practiced full-time by at least five people—in intervention areas of Comoé and Sourou were selected from those who had made at least one fishing expedition at the fishing site the day of the interview. Number of fishermen listed in fishing sites: N=538 (131 in Comoe; 407 in Sourou) Final sample size: 842 fishing trips.	Socio-economic characteristics of the fishermen; information on catches (number and types of fish caught), use of fish caught, labor employed; Information on the number of merchants and processors, and their equipment	Data **quality is acceptable.** Data have **low research value**: Fishing is not a part of the compact activities or the evaluation.	We propose neither a restricted access file nor a public use file, given low research value.
DA	Farmer training institutional survey	NORC and CERFODES, 2012 dry season (May)	The objective of the institutional survey data was to provide information on project activity related institutions for monitoring purposes. This data collection targeted institutions in communes and villages in the Sourou Valley and Comoé Basin related to livestock, fishing, forestry, and access to markets at the commune level. Data on topics such as water management, fee collection, infrastructure management, and satisfaction with the quality, availability, and management of water, were also collected from the old irrigated perimeters.	A total of 56 interviews were completed during the month of May 2012, of which 21 were in Comoé and 35 in Sourou. (see Table 2) 8 separate instruments.	Animal husbandry Fishing Forestry Market Water management Infrastructure Survey Producer Association Survey WUA Fee Survey	Data **quality is acceptable.** Data have **low research value**, given small sample sizes.	We propose neither a restricted access file nor a public use file, given low research value.

Activity	Evaluation and survey	Data collector and year	Objective and description of survey	Sampling and sample size	Modules / Content	Assessment	Recommendation
	Farmer training barymetric survey	NORC and CERFODES, dates unclear	The barymetric survey obtained data on cattle weight, milk production, and other related livestock data. Specifically: Quantity of milk produced per cow Livestock-type and number of animal Production and commercialization of animal products Livestock deaths Vaccinations In-Vitro fertilization Two rounds of data were collected: baseline and follow-up. Data collection dates are unclear in the documentation. Survey instrument only available for one round of data.	Baseline: N= 704 cattle- 494 in treatment area, 210 in control area from 153 households Follow-up: N= 565 cattle- 471 in treatment area, 94 in control area from 146 households	Cattle Weight Dataset- General information on cattle that belongs to household Household Dataset- General information on the types of animals that belong to household Milk Production Dataset- General information on cows owned by household	Data **quality is acceptable.** Data have **high research value**, given the lack of data-sets with livestock weight. **Full de-identification is feasible, but would be costly** for comprehensive public use file because participants with larger livestock herds could be easily identified. Consent language would allow data to be **shared confidentially with direct identifiers removed, as well as de-identified and shared publicly.**	Submit **restricted access file** to ICPSR Submit **key indicator public use file** with only region as geographic identifier (Table 5).
DA	Farmer training interim crop yield survey (IMPAQ)	CERFODES, 2013	Early interim yield information in both treatment and comparison areas of the ADP, using the crop yield square methodology.	Sample frame – Plots operated by the 2,164 households of the baseline survey: 7,241 total plots- 1,866 in treatment and 5,375 in comparison. Plots if growing focus crops (maize, sorghum, peanut, rice, millet, cowpea, sesame, groundnut): N=5,833 plots. (Due to late implementation, some squares were not laid as production already harvested) Yield squares laid (according to survey report): N=4,628. In data file: N=2494 plots, from 949 households	single module- yield squares	Data quality is **acceptable.** Data have **low** research value, because missing GIS coordinates. **Data de-identification is feasible, but would be costly** if linked to the baseline household survey. Consent language would allow data to be **shared confidentially with direct identifiers removed, as well as de-identified and shared publicly.**	We propose neither a restricted access file nor a public use file, given low research value.

Activity	Evaluation and survey	Data collector and year	Objective and description of survey	Sampling and sample size	Modules / Content	Assessment	Recommendation
	Farmer training supplemental household survey (IMPAQ)	CERFODES, 2013	Early interim survey that provides information related to the training and the support farmers received from AD10 and a short section on estimates of household production. Only collected in the treatment area.	Sample frame- 1,082 treatment households. In data file: N=949	Training; Household production	Data quality is **acceptable.** Data have **high** research value, in combination with ADP baseline data. **Data de-identification is feasible, but would be costly** if linked to the baseline household survey. Consent language would allow data to be **shared confidentially with direct identifiers removed.**	Submit as **restricted access file** to ICPSR with identifier links to farmer training baseline household data.

WMI= Water Management and Irrigation; DA=Diversified Agriculture; CERFODES=Centre d' Etudes, de Recherches et Formation pour le Développement Economique et Sociale; BERD= Bureau d'Etudes et de Recherche pour le Développement; PAP=persons affected by the project; ICPSR= Inter-university Consortium for Political and Social Research

Table 2. Overview of institutional survey contents and respondents

Institutional Survey	Questionnaire available	Number	Comoe	Sourou
Animal husbandry	Yes	4	1	3
Fishing	Yes	4	1	3
Forestry	Yes	4	1	3
Market	Yes	4	1	3
Water management	Yes	10	4	6
Infrastructure Survey		10	4	6
Producer Association Survey		10	4	6
WUA Fee Survey		10	4	6
Total		56	21	35

Table 3. Proposed indicators for the Di perimeter key indicator public use file (interim and final data collections)

Indicator	Sample restriction	Surveys
Survey Round	None	Interim and final surveys
Type of land (rice / polyculture)	None	Interim and final surveys
Yields by focus crop	None	Interim and final surveys
Total value of ag production	None	Interim and final surveys
Total ag profits, profits by focus crops	None	Interim and final surveys
Total cost of production	None	Interim and final surveys
Perception of land tenure security	PAPs only	Interim and final surveys
Perception of change in income and food security	PAPs only	Interim and final surveys
Sampling weight	None	Interim and final surveys
Household level information		

Note: We do not indicate the type of the beneficiary, land size or involvement in conflict as this facilitates re-identification. If we are able to obtain meaningful price information,

Table 4. Proposed indicators for the farmer training surveys key indicator public use file

Indicator	Sample restriction	Surveys
Region	Interim analysis sample	NORC baseline survey, Interim survey
Round	Interim analysis sample	NORC baseline survey, Interim survey
Type of irrigation*	Interim analysis sample	NORC baseline survey, Interim survey
Use of improved seeds, organic and inorganic fertilizer, and pesticide	Interim analysis sample	NORC baseline survey, Interim survey
ADP focus practices and adaptation of practices**	Interim analysis sample	Interim survey
Yields by focus crop**	Interim analysis sample	NORC baseline survey, Interim survey
Total ag profits**	Interim analysis sample	NORC baseline survey, Interim survey
Profits by focus crops**	Interim analysis sample	Interim survey
Total cost of production	Interim analysis sample	Interim survey
Total agricultural income	Interim analysis sample	NORC baseline survey, Interim survey
Household level information		

Note: We do not include area planted or total production by focus crop (which is one of the key research questions) since this would facilitate re-identification.

* Detailed categories may need to be combined, as for the Comoé areas there may be only a handful of observations with specific type of irrigation access.

** Only to the extent that these do not allow for re-identification. In particular, whether a farmer grows Soja might be an indicator that allows for easy re-identification.

Table 5. Proposed indicators for the barymetric surveys key indicator public use file

Indicator	Sample restriction	Surveys
Survey Round	None	Barymetric baseline and follow-up survey
Region	None	Barymetric baseline and follow-up survey
Village ID (masked)	None	Barymetric baseline and follow-up survey
Cattle breed	None	Barymetric baseline and follow-up survey
Cattle age	None	Barymetric baseline and follow-up survey
Cattle gender	None	Barymetric baseline and follow-up survey
Barymetric measurements	None	Barymetric baseline and follow-up survey
Milk production measurements	None	Barymetric baseline and follow-up survey
Daily milk production measurements last month	None	Barymetric baseline and follow-up survey
ADP training receipt	None	Barymetric follow-up survey
Livestock level information		

Note: We include masked numeric village ID to allow for intra-cluster correlation computation.