Relationship of Weather and Maize Yields in Kenya

Monika Novackova, Pedram Rowhani, Martin Todd, Annemie Maertens

Department of Geography, University of Sussex, Falmer, UK

October 2018

Abstract

We explore an unprecedented dataset of almost 6,000 observations to identify main predictors of climate knowledge, climate risk perception and willingness to pay for climate change mitigation. Among nearly 70 potential explanatory variables we detect the most important ones using multisplit lasso estimator. Importantly, we test significance of individuals' preferences about time, risk and equity. Our study is innovative as these behavioural characteristics were recorded by including experimental methods into a live sample survey. This unique way of data collection combines advantages of survey and experiments. The most important predictors of environmental attitudes are numeracy, cognitive ability, ideological world-view and inequity aversion.

JEL classification: Q54, Q58, D80

Keywords: Climate change, climate knowledge, climate policy, lasso, risk perception, willingness to pay

1 Introduction

Findings:

- OND last year dry spell, max rain very important for Maize, but cumulative precipitation for the same period not so important
- Mar-Sept last year temperature very important for maize yields
- SD temperature last year positive and significant
- dry spell 20 MAM last year important (but not dry spell MAM10)
- interesting. Precipitation 2 months MAM last year very significant and positive
- mean temp last year negative and significant, hill shaped

New findings:

- The yields seem to be more responsive to weather on west than on east

2 Methodology

Prior empirical studies detected large number of miscellaneous predictors of climate change knowledge and concerns (e.g. **????**). There is, however, a lack of consensus about which

are the most important ones. Since our dataset includes almost 70 potential predictors, we decided to start with an explanatory regression analysis using a model selection estimator.

Stepwise-like procedures were found to be problematic as it was shown that large portion of selected variables is often noise and the adjusted R^2 is biased upwards (**?**). There are also other problems with these methods. For example, a forward stepwise regression selects in each step the predictor having largest absolute correlation with the response y, say x_{j1}. Then a simple linear regression of y on x_{j1} is performed and a residual vector from this regression is considering to be the new response variable. Then the procedure is repeated and we eventually end up with a set of selected predictors $x_{j1}, x_{j2}, ..., x_{jk}$ after k steps. This method can, however, eliminate a good predictor in second step if it happens to be correlated with x_{j1}. Furthermore, these methods frequently fail to identify the correct data generating process, even in large samples (**?**). A possible alternative is the best subset selection approach. Given a collection of possible predictors, the best subset approach compares all possible subsets of predictors based on some well-defined objective criterion, usually having the largest adjusted R^2. However, besides being excessively computationally demanding, also this method often fails to identify the true predictors (**?**). On the other hand, sparse estimators such as lasso (**?**) are usually more stable than stepwise procedures and they are commonly better in prediction accuracy (**?**). Because lasso has been shown to be very powerful for high-dimensional variable selection in general (**?**), we opt for this estimator.

Using the same notation as **?**, our dependent variable is $Y \in \mathbb{R}$ and our vector of explanatory variables is $X \in \mathbb{R}^p$. We assume that the relationship between them can be approximated by a linear regression model $E(Y|X = x) = \beta_0 + x^T\beta$. Lasso estimator selects the predictors by setting some of the coefficients β_j to be equal to zero.

We consider four distinct models for the four response variables and one additional model

as a robustness test. The dependent variables are: (i) Knowledge about climate change (ii) Perceived seriousness of climate change (iii) Perception of effects of climate change policy relatively to effects of climate change and (iv) WTP for climate change mitigation, which we measure by preferred tax rates on gas and electricity. We also estimate an additional model for petrol duty as a robustness test for the WTP model. How we measure the dependent variables is described in Section 3.1. The potential predictors included in x, which are not the behavioral variables and which were not selected into any model by multisplit lasso are listed in Tables A1 and A2 in Appendix 3. How we measure the behavioural variables is discussed in Section 3.2 and their descriptive statistics are summarised in Table A3 in Appendix 3 with the exemption of inequity aversion as this variable is considered as categorical and its frequencies are summarised in Table **??** in Appendix 3. The predictors, which were selected into some model can be found in a table of estimates of the relevant models and their descriptive statistics or frequencies are summarised in Tables A3, **??**, **??**, and **??** in Appendix 3.

The estimation function can be written as (**?**):

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{(p+1)}} \boldsymbol{R}_\lambda(\beta_0,\beta) \quad = \quad \min_{(\beta_0,\beta)\in\mathbb{R}^{(p+1)}} \left[\frac{1}{2N}\sum_{i=1}^{N}(y_i-\beta_0-x_i^\intercal\beta)^2+\lambda\sum_{j=1}^{p}(|\beta_j|)\right], \qquad (1)$$

where y_i is the value of one of our four dependent variables for an individual i, x_i includes potential predictors listed in Tables A1 to **??** in Appendix 3, N is the number of observations and $\lambda \geq 0$ is the penalty parameter. Without loss of generality, we assume that the potential predictors in (1) are standardized: $\sum_{i=1}^{N} x_{ij} = 0$, $\frac{1}{N}\sum_{i=1}^{N} x_{ij}^2 = 1$, for $j = 1, ..., p$. [1]

[1]Both x_{ij} and y_j are standardized automatically in the implementation of the algorithm we use. However, the estimated coefficients are always returned and presented on the original scale.

In line with common practice, we compute estimator (1) for a series of λ and then we choose a preferred value of λ using cross-validation (**?**). In particular, we use a sequence of 100 values of λ and 10-fold cross validation.[2] We opt for the value of λ which is recommended by **?** and it is probably the most common choice. More specifically, we use the largest value of λ such that the mean cross-validated error (CVM) is still within one standard error of its minimum.[3]

Determining significance levels is problematic with lasso. Classical p-values are not valid and there is no simple approximation. Therefore, we adopt a concept of **?**, who introduce an approach based on multiple random splits of data, repeated estimation and aggregated inference. In particular, **?** build on the proposal of **?**, who suggest to split the dataset randomly into two subsets. One of the subsets is used for variable selection via lasso and the other one is for estimating OLS with the predictors selected by lasso and calculating their p-values in a usual way. This procedure allows asymptotic error control under minimal conditions. The problem is that the results depend on a one-time arbitrary split and they are therefore irreproducible. **?** further develop the single-split method. They suggest to split the sample repeatedly, obtain a set of p-values for each split and then aggregate them. In each split, the p-values of the variables which are not selected are considered to be equal to one and the p-values of the selected variables are multiplied by the number of variables selected in the current split. If a p-value multiplied by the number of selected variables happens to be larger than one, it is considered to be equal to one. Let's assume that we have $h = 1, ..., H$ splits. A p-value for predictor j obtained in split h adjusted as described above will be further denoted $P_j^{(h)}$. **?** suggest to aggregate the adjusted p-values using

[2]For estimation of lasso (1) we use function cv.glmnet in the **R** programming system (**?**) and we use default settings and values of arguments, unless otherwise stated.

[3]In case of WTP we use the value of λ which minimises the CVM. This value is also suggested by **?**. The only difference from the model estimated using the one standard error based λ is that for the latter, a dummy variable for male becomes significant and gets into the model. The effect of male is positive and this contradicts predominant conclusions in previous relevant literature (e.g. **????**).

quantiles. In particular, a suitable aggregated p-value is defined for any predictor j and for any fixed $0 < \gamma < 1$ as

$$Q_j(\gamma) = min\left\{1, q_\gamma(\{P_j^{(h)}/\gamma; h = 1, ..., H\})\right\}, \tag{2}$$

where and $q_\gamma(\cdot)$ is the (empirical) γ-quantile function. We will further refer to this procedure as a multisplit lasso.

? show that for any predefined value of $\gamma \in (0, 1)$, the p-values defined in (2) can be used for control of family-wise error rate[4] and also for regulation of false discovery rate.[5] Moreover, the multisplit method improves the power of estimates.

3 Data

All data used in this study except of predicted income and population density, which we use in robustness tests, were collected in the survey conducted by **?**.

In Section **??** we use an alternative measure of income as a robustness test. In particular, this estimated income is obtained from a regression model based on data from Annual Survey of Hours and Earnings (ASHE). More specifically, the predicted income is based on age, gender, occupation, sector and education.

We use two measures of population density, in particular average density per Lower Layer Super Output Areas (LSOA) estimated by the Office for National Statistics for year 2015 and average density for Local Authority Districts (LAD) obtained from the 2011 Census.

The online survey (**?**) ran from 9 September to 14 October 2015 and 6,000 respondents

[4]Probability of making at least one incorrect rejection of a true null hypothesis (type 1 error).

[5]Expected proportion of incorrect rejections of a true null hypothesis (type 1 errors). False discovery rate controlling procedures are less stringent than family-wise error rate controlling methods.

were selected to answer the questionnaire which included the climate change domain.[6] Descriptive statistics, methodology of the survey, the survey itself and a detailed description of its administration can be found in **?**.

The survey is reasonably geogrephically representative taking into account population density in the UK (**?**).[7] As the survey was conducted online, the initial sample is representative for UK adults with internet access rather than for the entire UK population.

In Table 1 we compare distribution of our sample over sex and age with the distribution of the UK population. The age data are only available as a categorical variable in our survey. As we can see in Table 1, the youngest category is slightly over-sampled while the two categories of the highest age are slightly under-sampled, probably because the survey was conducted online. Otherwise the distributions are very comparable.

[6] We had to exclude some observations from various parts of analysis as they included missing values for some important variables. However, we have at least 5500 observations for each model.

[7] For map with location of respondents see Figure 1 in **?**

Table 1: ***Sex and age distribution***
of the sample and the population

	Sample		UK population[a]	
Age range	**Male**	**Female**	**Male**	**Female**
$18-24$	9.8%	9.4%	6.2%	6.0%
$25-34$	10.0%	10.3%	9.0%	9.1%
$35-44$	7.8%	8.3%	8.6%	8.8%
$45-54$	8.1%	9.4%	9.3%	9.6%
$55-64$	7.4%	8.4%	7.5%	7.8%
$65-74$	4.1%	4.8%	6.2%	6.7%
$75-80$	0.1%	0.1%	2.4%	2.8%

[a] Population data are from the Office of National Statistics, Population Estimates of UK, England and Wales, Scotland and Northern Ireland Mid 2014, Table MYE2.

Although the survey questionnaire was designed such that more difficult questions were at different pages, we observe that most respondents who did not finish the survey dropped out on pages with more difficult questions. Hence, the final sample is biased towards those who are not afraid of hard questions (**?**).[8]

In the rest of this section we focus on how we obtained the data for our climate (dependent) variables and the behavioural characteristics..

[8]One way how to deal with sample selection is to use sampling weights. We, however decided not use weights given the modest nature of our bias. Weighting usually increases standard errors and leads to less precise estimates and there is lack of consensus on whether or not to use weights in regression methods (**???**). **?** for example recommend not to use weights if they are solely a function of independent variables.

3.1 Climate variables

Descriptive statistics of our climate variables are summarised in Table 2.

Table 2: ***Dependent variables:*** *Descriptive statistics*

Variable:	**Mean**	**St. dev.**	**Min**	**Max**
Climate change knowledge	3.851	1.266	1	8
Climate change seriousness perception	6.622	2.249	0	10
Climate versus policy effects perception	5.370	2.315	0	10
WTP - gas and electricity tax (£ per year)	123.900	105.459	0	500
WTP - duty on transport fuel (pence per year)	20.530	22.518	0	100

It was previously shown, that questions which are intended to measure climate science comprehension often measure who people are rather than what they know about climate change as the strongest predictor is often respondents' ideology and cultural and political world-view (**???**). To avoid picking of effect of cultural or political world-view instead of climate knowledge, we use questions from the OCSI instrument developed by **?** as a measure of climate knowledge. **?** shows that these questions are indeed a measure of climate science comprehension rather than an indicator of who one is. The values of climate knowledge are integers from 0 to 8 and they stand for counts of correctly answered questions about climate change (**?**). An example of one of the 8 climate questions is: 'Climate scientists believe that if the North Pole icecap melted as a result of human-caused global warming, global sea levels would rise. Is this statement true or false?' The list of all climate questions can be found in Appendix 1. The relative frequencies of counts of the correctly answered questions are summarised in Table 3.

To investigate opinions about seriousness of climate change, the respondents were asked the following question: 'How serious a problem do you think climate change is at this moment?' Using an interactive slider, the respondents answered an integer value between 0 and 10 where min = 0 and max = 10 (as it was noted just below the slider). In a similar way, the respondents were asked if they feel to be more affected by climate change or by climate policy. The wording of the question was: 'Which affects you and your way of life more, climate change or policies to reduce greenhouse gas emissions?' Again, the respondents provided answers on an integer scale from 0 (climate policy) to 10 (climate change) using a slider. Relative frequencies of climate seriousness perception and climate versus policy perception are summarised in Table 3.

Table 3: ***Dependent variables:*** *Relative frequencies (%)*

Variable:	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Climate knowledge	0.0	1.7	11.4	30.4	25.4	20.9	8.6	1.6	0.1	N/A	N/A
Climate seriousness perception	3.3	2.8	5.4	8.1	8.9	27.2	14.0	12.8	8.3	4.1	5.0
Climate vs. policy perception[a]	2.1	1.5	2.5	3.8	4.6	9.8	18.5	21.7	16.7	8.6	10.4

Notes: Total number of observations: 5749

a Higher number means greater concern about climate change, lesser concern about climate policy.

Regarding the preferred gas and electricity tax rates, the respondents were first asked how much the current tax was. In particular, the question was as follows: 'The average household pays £1,369 per year for gas and electricity. Government intervention has raised

the price to encourage people to use less and so reduce greenhouse house gas emissions. How much of that £1,369 is for climate policy?' They indicated the response on a slider with a minimum of −50 and a maximum of 500. We include this variable on right hand site as a robustness test (see Table **??**). We refer to it as 'How much is tax gas and electricity'. After this, the respondents were told the correct answer and they were asked about they preferred tax rates: 'Actually, climate policy adds about £89 per year to the gas and electricity bill of the average household. How much do you think climate policy should add to this bill?' The respondents expressed their opinion on a slider from 0 to 500. The answer to this question is the dependent variable which we refer to as 'WTP - gas and electricity' and we use it as a proxy for WTP for climate change mitigation. Analogously, we inquired about the fuel duty. The only difference is that the slider for the actual fuel duty is limited from 0 to 60 and the one for the preferred fuel duty is from 0 to 100 as the actual fuel duty is 3 pence per litre. Descriptive statistics of the respondents' estimates of actual tax rates can be found in Table A3 in Appendix 3 and the descriptive statistics of the preferred tax rates are in Table 2.

3.2 Behavioural variables

To estimate the social value orientation, respondents played six dictator games with the same questions as in **?**. The ring measure of social value orientation which we use in our models is defined as

$$R = \arctan \frac{\sum_{i=1}^{N} P_O - 50N}{\sum_{i=1}^{N} P_S - 50N}, \tag{3}$$

4 Results and discussion

In this section we describe our results and discuss their interpretation.

In the tables which summarise the estimates of lasso below, p-values of some of the explanatory variables are equal to one. These variables were not selected by the lasso in most of the sample splits. They are, however, included in the tables because they represent either a category of a nominal variable whose other category was selected by the lasso or a linear term of a variable whose quadratic term was selected by the lasso.

4.1 Climate change knowledge

Table 4: ***Relationship of weather and maize yield:*** *Mixed models*

	Estimated coefficients and p-values		
Fixed effects	**Linear - unscaled**	**Linear-scaled**	**Log-linear**
Intercept	−0.059	0.646	
Prec. cum. MAM + OND lag 1, east	0.054	0.041*	
Prec. cum. MAM lag 1, west	0.025	0.477	
Temp. avg. Mar. - Sep. lag 1, east	−0.045	0.244	
Temp. avg. Mar. - Sep. lag 1, west	−0.045	0.244	
Prec. max OND, east	0.0.133	0.013*	
Prec. max OND, west	0.144	0.004**	
Temp. sd. Oct. - Mar. lag 1, east	0.049	0.199	
Temp. sd. Oct. - Mar. lag 1, west	0.225	4×10^{-12} ***	
Observations:	584		

Notes: • $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

a comment

a another comment

We find the positive and strongly significant effect of dummy variable for males quite peculiar. Previous research shows mixed evidence about effects of gender on climate knowledge and comprehension of science in general. For example, **?** finds that women demonstrate higher level of scientific knowledge of climate change. On the other hand, **?** shows that men exhibit significantly higher level of scientific knowledge than women, even if controlling for a number of background variables. We perform additional tests to verify whether the positive effect of gender can be a result of sample selection. The tests include proportion tests, model with interactions as additional explanatory variables and a Heckman selection model. We discuss the results in detail in Appendix 2. Based on the outcomes, we conclude that the results are not driven by sample selection.

A possible explanation why our measure of climate knowledge is significantly higher for men is that the climate knowledge test that we use in this study was developed by a man (**?**), therefore it may be the case that these particular questions are naturally more comprehensible for men. The only way how to test this would be to let a woman design another set of climate knowledge questions and then conduct a survey which would include these woman-designed climate questions. This is, however, beyond the scope of this study.

To sum up, we find that gender and cognitive ability are significant predictors of climate knowledge. Climate knowledge increases with higher numeracy which is consistent with **?**, who finds the climate knowledge measure to be positively correlated with ordinary science intelligence. Although various measures of climate knowledge were previously find to be correlated with social ideology or partisan identity (**???**), our measures of ideology, cultural world-view or their interactions were not chosen as predictors of climate knowledge by the lasso. This is also consistent with **?**.

4.2 Climate change risk perception

In this section we discuss our estimates of the models which explain individuals' perception of climate change risk. We focus on two measures of climate risk perception, in particular climate change seriousness perception and climate versus policy perception. We present the results of lasso and jackknife OLS with the climate seriousness perception as dependent variable in Table 5. Three predictors were selected, in particular gender, climate knowledge, and degree of agreement with redistribution of income by government. In this case, the effect of being male is negative. This is mostly consistent with results of previous research which typically finds women to take climate risk more seriously than men (**???**). As we can see in Table 5, degree of agreement with income redistribution affects climate change seriousness perception positively as the base category is 'Strongly disagree'. This is in agreement with previous literature as we consider the degree of agreement with income redistribution as an indicator of political and ideological world-view, which was found to be significantly correlated with climate concern by large number of previous studies (e.g. **???**).

We will comment on the significant effects of climate knowledge at the end of Section 4.2.

Table 5: ***Climate change seriousness perception:*** *Multisplit lasso and jackknife OLS*

	Multisplit lasso		Jackknife OLS		
Variable	**Aggregated adj. *p*-value**		**Aggregated coefficient**	**Aggregated adj. *p*-value**	
Gender = male	0.0002	***	−0.3658	4.45×10^{-6}	***
Climate knowledge	1.0000		0.1380	1.0000	
Climate knowledge - squared	$< 2.00 \times 10^{-8}$	***	−0.0548	0.0209	*
Redistribution of income: disagree[a]	1.0000		0.1819	1.0000	
Redistribution of income: neutral[a]	1.0000		0.2789	0.8251	
Redistribution of income: agree[a]	$< 2.00 \times 10^{-8}$	***	0.8343	8.58×10^{-8}	***
Redistribution of income: strongly agree[a]	$< 2.00 \times 10^{-8}$	***	1.0828	$< 2.00 \times 10^{-8}$	***
Observations:			5749		

Notes: • $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

For the significant predictors, the signs of the coefficients of the multisplit lasso are the same as those of the jackknife OLS and also size of most of the coefficients is very comparable for these two models.

a Degree of agreement with the following statement: 'Government should redistribute income from the better off to those who are less well off.' The base category is 'Strongly disagree'.

Appendix 3 Tables

Table A1: *List of considered (but not selected) predictors in multisplit lasso*

Variable	**Description**
Religion	11 categories including atheist, no religion and prefer not to say
Race	8 categories including prefer not to answer
Length in UK	Question: *How long have you been living in the UK?* Response = 5 categories: All life ,more than 10 years, $5-10$ years, $1-5$ years, less than 1 year
Occupation	14 categories
Sector	18 categories
Operating system	7 categories
Social value orientation	Response = 4 categories: altruist, prosocial, individualist, competitive
Discount rate 0 vs. 5	Annual, %, invest now for five years from now
Discount rate 1 vs. 2	Annual, %, invest a year from now for two years from now
Discount rate 1 vs. 6	Annual, %, invest a year from now for six years from now
Degree of present bias	Continuous, preferences on time
Degree of hyperbolicity	Continuous, preferences on time
Annual discount rate	Continuous, preferences on time
Subsistence income (reserve)	Continuous, **???**
Altruist	Dummy (0/1)
Prosocial	Dummy (0/1)
Individualist	Dummy (0/1)
Competitive	Dummy (0/1)
Egalitarian	Dummy (0/1)
Ineqaverse	Dummy (0/1)
Longitude	Longitude of survey response. Degrees
Latitude	Latitude of survey response. Degrees
Letter	First letter of surname, A=1,B=2,...
Siblings	Number of siblings
Older	Number of older siblings
Children	Number of children
Grandchildren	Number of grandchildren

Note: Variables in this table were not selected by multisplit lasso into any model.

(*continued*)

Table A2: *List of considered (but not selected) predictors in multisplit lasso*

Variable	**Description**
Handedness	0=right, 1=left
Time	Time taken to complete survey, in minutes
Hour	Hour of survey, 24 categories
Day of week	7 categories
Day of the month	Day of survey, $1-31$
Fair share	*Ordinary working people do not get their fair share of the nation's wealth.* Degree of agreement with the statement above, 5 categories
Hard work	Question: *How important is hard work for getting ahead in life?* Response = 5 categories, degree of agreement
Better off parents	Question: *Compared with your parents when they were about your age, are you better or worse in your income and standard of living generally?* Response = 5 categories (degree of agreement) and *Don't know*
Better off children	Q: *Compared with you, do you think that your children, when they reach your age, will be better or worse in their income and standard of living generally?*Answer =5 categories (degree of agreement) and *Don't know*
Always up	Dummy (0/1), Children better off me and me better off parents
Always down	Dummy (0/1), Parents better off me and me better off children
Up then down	Dummy (0/1), Me better off parents and me better off children
Down then up	Dummy (0/1), Parents better off me and children better off me
Financial literacy	3 financial problems, no. of correct answers, **?**
Understands portfolio	Dummy (0/1), 1 = understands
Incoherent dr.	Dummy (0/1), Incoherent answers between investments (0 = coherent)
Primed attitudes	1 = priming questions about time, risk, social were asked, 0 = not
Prime climate	0 = shown picture of polar bear on melting ice (negative), 1 = shown picture of people enjoying beach (positive)
Prime pension	0 = picture of troubled old man, 1 = picture of happy old man
Prime school	0 = picture of unruly kids, 1 = picture of well-behaved kids
Prime NHS	0 = picture NHS in crisis, 1 = picture love NHS
Female × handed	Interaction female and handedness
Female × children	Interaction female and number of children
Age × children	Interaction age and number of children

Note: Variables in this table were not selected by multisplit lasso into any model.

(*continued*)

Table A3: ***Descriptive statistics:*** *Continuous variables*

Variable:	**Mean**	**St. dev.**	**Min**	**Max**
Income - predicted (£ per year)	27729	11719.89	3611	58326
Net assets - total assets minus total debts (£)	152542	223612.90	−400000	2500000
Population (per Km^2, LSOA[a] level)	3336	2975.38	7	25280
Population (per Km^2, LAD[b] level)	3193	3164.75	10	13870
How much is tax gas and electricity (£/yr.)	144.90	111.94	−50	500
How much is duty transport fuel (pence/yr.)	25.18	13.68	0	60
Behavioural variables				
Social value orientation (ring measure)	26.28	15.52	−16.26	83.93
Annual discount rate,%, invest now for a year from now[c]	148.7	181.81	1	500
Risk aversion - estimated median of quadratic utility function	0.33	0.01	0.29	0.38
Risk aversion - estimated median of log utility function	1.81	1.08	0.67	4.33
Risk aversion - estimated median of power utility function	0.42	0.07	0.33	0.57
Risk aversion - estimated mean of power utility function	0.74	0.26	0.33	1.07

Notes: Total number of observations: 8541

a Lower Layer Super Output Area

b Local Authority District

c This variable is called *Discount rate year from now* in the tables with regression estimates