Chapter 2: A new coffee production database

Jeffrey Sachs, James Rising,
Tim Foreman, John Simmons, Manuel Brahm
The Earth Institute
Columbia University

October 1, 2015

Prepared for the Global Coffee Forum
Milan, Italy

Part of
The impacts of climate change on coffee: trouble brewing
`http://eicoffee.net/`

In this report, we identify the relationships between climate and coffee through careful, empirically-grounded methods. Identifying the locations of coffee production is essential for understanding how coffee is already interacting with the climate and how it will respond to climate change. High resolution weather station and gridded weather data are readily available, to identify regions subjected to high temperatures, frosts, precipitation, and humidity, but their impacts is most clear when the weather and coffee data are closely aligned in space. Similarly, climate change suitability maps are more useful when compared to high-resolution information about the current location of coffee growing areas.

Applying robust spatial methods requires a new global database of coffee production. We develop an initial version of this database, combining existing records of coffee production with geospatial maps of coffee producing areas. However, much work remains to be done. Over 50 countries produce coffee, but information stored in the database for most of these countries is sparse.

Many groups can contribute to the database as we have created it. As part of the database, we have developed standardized formats and a clear flow for combining multiple kinds of spatial and temporal data. Organizations involved in collecting and organizing coffee production data can contribute to a collaborative process, which together will provided better data for a large range of research activities.

Previous global datasets only provide coarse information on coffee production. The most reliable global information on coffee production, at the ICO, the FAO, and the USDA Foreign Agricultural Service, is only available on a per-country basis. CIAT has constructed a database of information on coffee farms (see figure 1), but our analysis requires production information as it changes over time. To our knowledge, there is no previously existing global dataset of coffee producing regions at the subnational level.

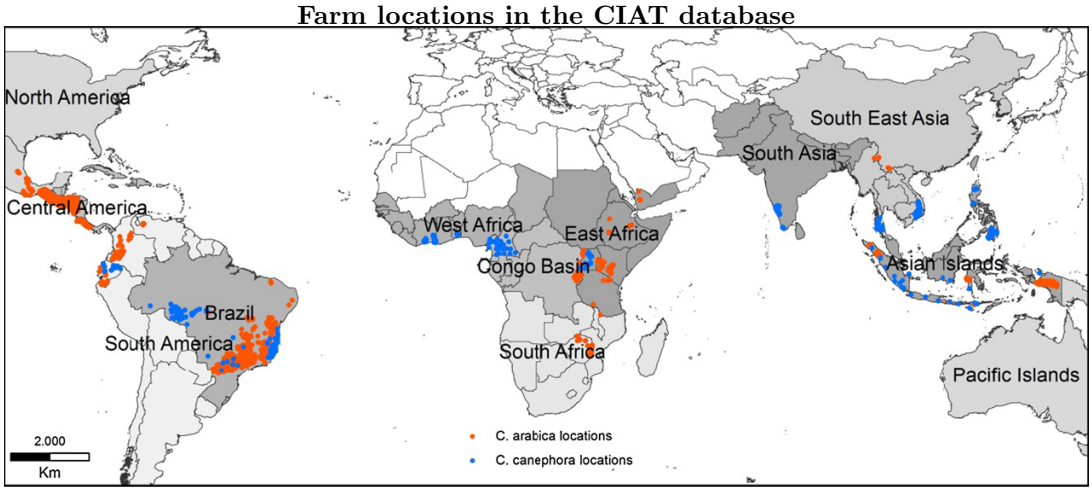


Figure 1: Arabica and Robusta coffee producing farms in the CIAT database.

Monfreda et al. (2008) provide an approximate geographic distribution for coffee, by first identifying global cropland at high resolution (5'), and then using country-specific databases of coffee production, where available, to refine the areas. The quality of the resulting production areas varies widely by country, as shown in figure 2. For most countries, only country-level production is available. Four countries have county-level data on coffee production, and 13 others have state-level production information.

We have combined the information from the FAO, the UDDA, the CIAT farm map, and Monfreda et al. with detailed maps of coffee production regions for 8 major coffee producing countries. These countries produce roughly 85% of the world's coffee. These high-resolution country-specific maps allow us to assess the characteristics of coffee production at much greater detail, both in terms of climate variables

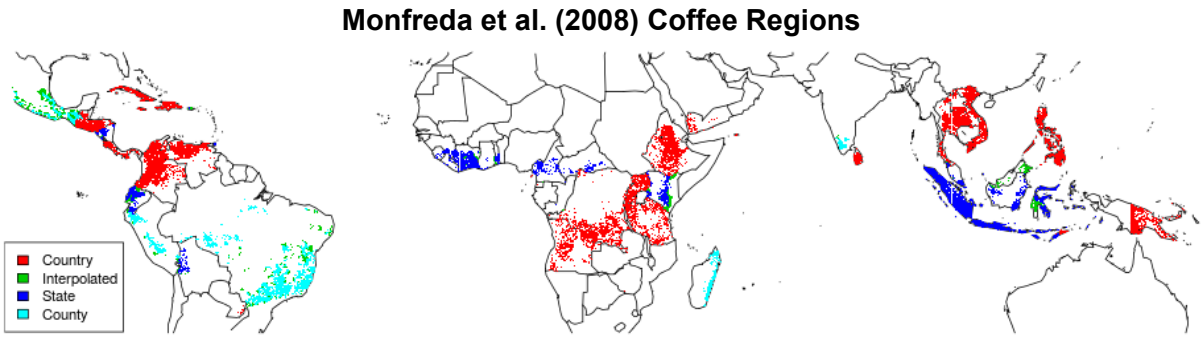


Figure 2: Quality of geospatial production data for coffee, from Monfreda et al. (2008). Global agricultural areas are intersected with country-specific datasets to create a global map of coffee production areas.

such as temperature and precipitation, and, in chapter **??**, geographical variables, such as latitude and elevation. Combining these factors with data on yields in these areas allows us to estimate the effects of weather patterns on coffee growing, as well as study the range of climates where coffee can be grown successfully.

A summary of this information is provided in table 1 and some production area maps are included in appendix 0.3, along with a description of the process for combining these maps.

Country	Production	Coverage	Resolution	Source
Brazil	2,720 kt	country-wide	municipality (5503)	IBGE
Vietnam	1,650 kt	country-wide	raster image	Cafecontrol
Colombia	696 kt	country-wide	raster image	Oficina de E. y P. Básicos Cafeteros
Indonesia	411 kt	6 regions	raster image	Schroth (2014)
Ethiopia	390 kt	country-wide	raster image	GAIN (2013)
India	300 kt	country-wide	state (13)	Coffee Board (India Gov.)
Mexico	270 kt	country-wide	raster image	**?**
Guatemala	240 kt	country-wide	vector layers	MFEWS
El Salvador	82 kt	country-wide	raster image	Poyecto Programa Ambiental
Nicaragua	78 kt	country-wide	raster image	MFEWS
Tanzania	50 kt	country-wide	raster image	Caparo et al. (2015)
Haiti	21 kt	country-wide	vector image	Coffee Supply Chain Risk Ass. Miss.
Rwanda	21 kt	country-wide	points	Nzeyimana (2014)
Yemen	14 kt	country-wide	raster image	**?**
Total	7,766 kt	global	country	ICO, FAO, USDA FAS
		global	inferred raster	Monfreda et al. (2008)
		global	points	Bunn et al. (2015), CIAT

Table 1: Sources of spatial coffee production, from academic literature, government agencies, and NGO reports. Average production values are taken over the past decade. The resolution is listed as either a reporting level (municipality, state, country), as a graphical map (raster image, vector layers), a gridded analysis (inferred raster), or individual points (points).

Similarly, yield data for coffee at a finer resolution than the country level helps identify more closely the impact of existing climate dynamics on yield. Table 2 summarizes the data collected on yields, production, area planted and harvested, and fertilizer use. Yields are typically reported as the quantity of coffee produced in an area (which we standardize to metric tonnes), divided by the area undergoing

harvest which we report in hectares). A better measure of yield would divide the number of hectares that was planted, as measured two to three years before the harvest to capture the potential loss of plants before and after maturity.

Country	Variables	Time	Space	Organization
India	Planted, Produced, Yield	32 years (1951 - 2013)	15 growing regions	Knoema
Brazil	Harvested, Produced	yearly (1990 - 2012)	5624 municipalities	IBGE
Indonesia	Area, Produced, Yield	2011	20 districts (Kecamatan)	Dinas Pertanian
Rwanda	Area, Agroforested, Yield	2005	10 growing regions	NAEB
Vietnam	Area	2012, 2013	11 provinces + other	GAIN
Brazil	Fertilizer use	2002	5 regions	FAO
global	Harvested, Produced, Yield	1961 - 2012	86 countries	FAO
global	Produced (by variety), Stocks, Export, Consumption	1960 - 2013	79 countries	USDA FAS
global	Fertilizer use	1995 - 2002	24 countries	FertiStats

Table 2: Sources of global and sub-country data on coffee yields, total production, and planted and harvested area.

0.1 Confidence maps

The first result of the database is its own measure of confidence in the geographic data across the globe. The confidence maps reflect the combined amount of information available, across the multiple map inputs. Each contributing map is assigned its own confidence, with maps of global harvest having low or medium confidence and maps detailing a given country with high confidence. Where multiple input maps corroborate each other, the confidence increases (see appendix B.1). In figure 3, dark green represents low confidence, and yellow and tan colors represent high confidence. The band of lighter green in the middle shows the overlap between maps from Thurston et al. (2013), Monfreda et al. (2008), and Bunn et al. (2015).

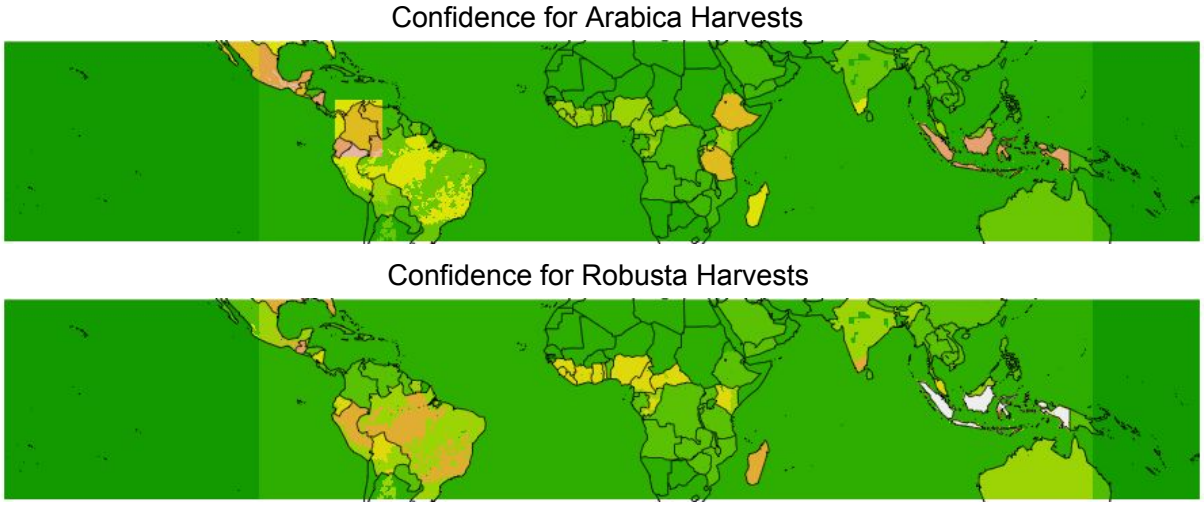


Figure 3: Database geospatial harvest confidence, based on the amount and scale of data available.

0.2 Harvest maps

The harvest maps are the main output of the spatial portion of coffee database. For each month, these combine country-specific information (some of which specifies harvest months as they differ across the country), with global harvesting regions applied to a calendar of harvest months from Sweet Maria (2015). Some country calendars are unavailable, so these harvested regions show throughout the year. The added weight of these multiple instances will be handled next.

A visual representation of the coffee database is shown in figure 6. We use average country-wide total harvest areas from FAO to translate harvest patterns into a description of the portion land area that is harvested. In areas where only country-level data is available, such as China, the entire country is shown as having a uniform low harvest. Where detailed coffee production data is available, such as in Brazil, the intensely cultivated areas are distinguished from those where coffee is absent. The values shown in the figure are used to combine weather information when estimating the effects of weather revealed in country-wide production records.

A visual representation of the coffee database is shown in figure 6. We use average country-wide total harvest areas from FAO to translate harvest patterns into a description of the portion land area that is harvested. In areas where only country-level data is available, such as China, the entire country is shown as having a uniform low harvest. Where detailed coffee production data is available, such as in Brazil, the intensely cultivated areas are distinguished from those where coffee is absent. The values shown in the figure are used to combine weather information when estimating the effects of weather revealed in country-wide production records.

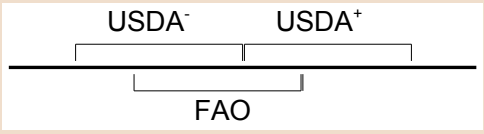
0.3 Time series data

Both FAO and the USDA Foreign Agricultural Service report production information for coffee, but the information they provide is quite different. FAO reports total production and harvested area, for all varieties of coffee combined, with a total of 4242 observations. The USDA reports only production information, but divides it out by Arabica and Robusta production, with 3211 observations per variety. The number of countries included also varies by year (see figure 7).

A second complication arises from the definition of the reported year. FAO reports production for calendar years, while USDA reports it for market years which vary by country. This can be an opportunity, allowing us to determine more precisely when production occurs. For example, in Brazil, coffee is harvested mainly between May and September. However, the USDA market year for Brazil is from July to June. So, discrepancies between the FAO and USDA production totals allow us to distinguish, approximately, between the share of production before and after July, the start of the market year cycle.

Calculating intra-year production

The diagram below shows how the USDA and FAO calendars align. The actual division is different for each country, depending on the start of the USDA market year.



We divide each USDA value into "left" and "right" parts, with $USDA^L = \alpha USDA$ and $USDA^R = (1-\alpha)USDA$, where the coefficient α is unknown. Further, we know from the diagram that

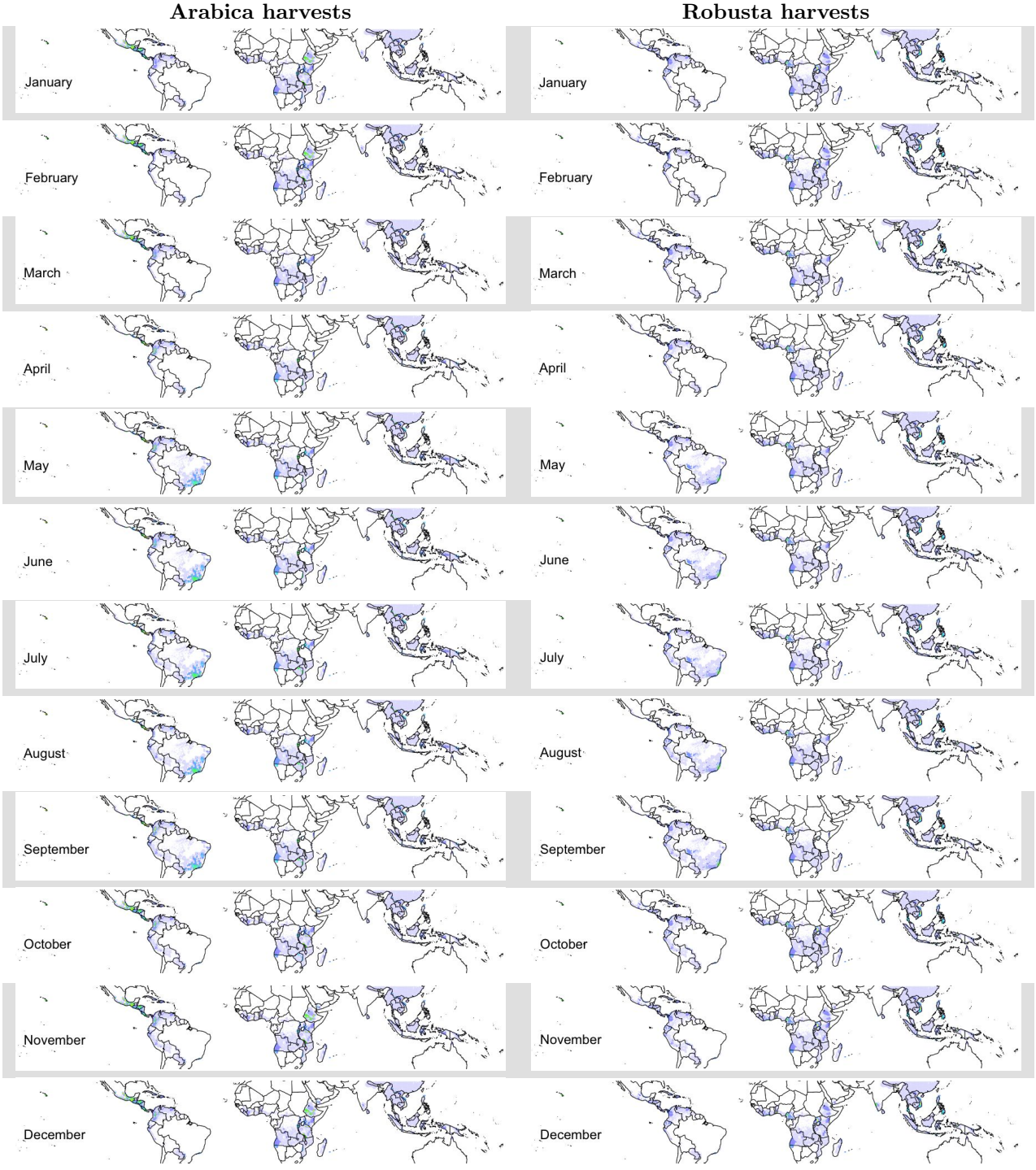


Figure 4: Harvest maps for Arabica and Robusta varieties, during each month. Darker colors represent higher levels of evidence that these regions are undergoing harvest in the given month.

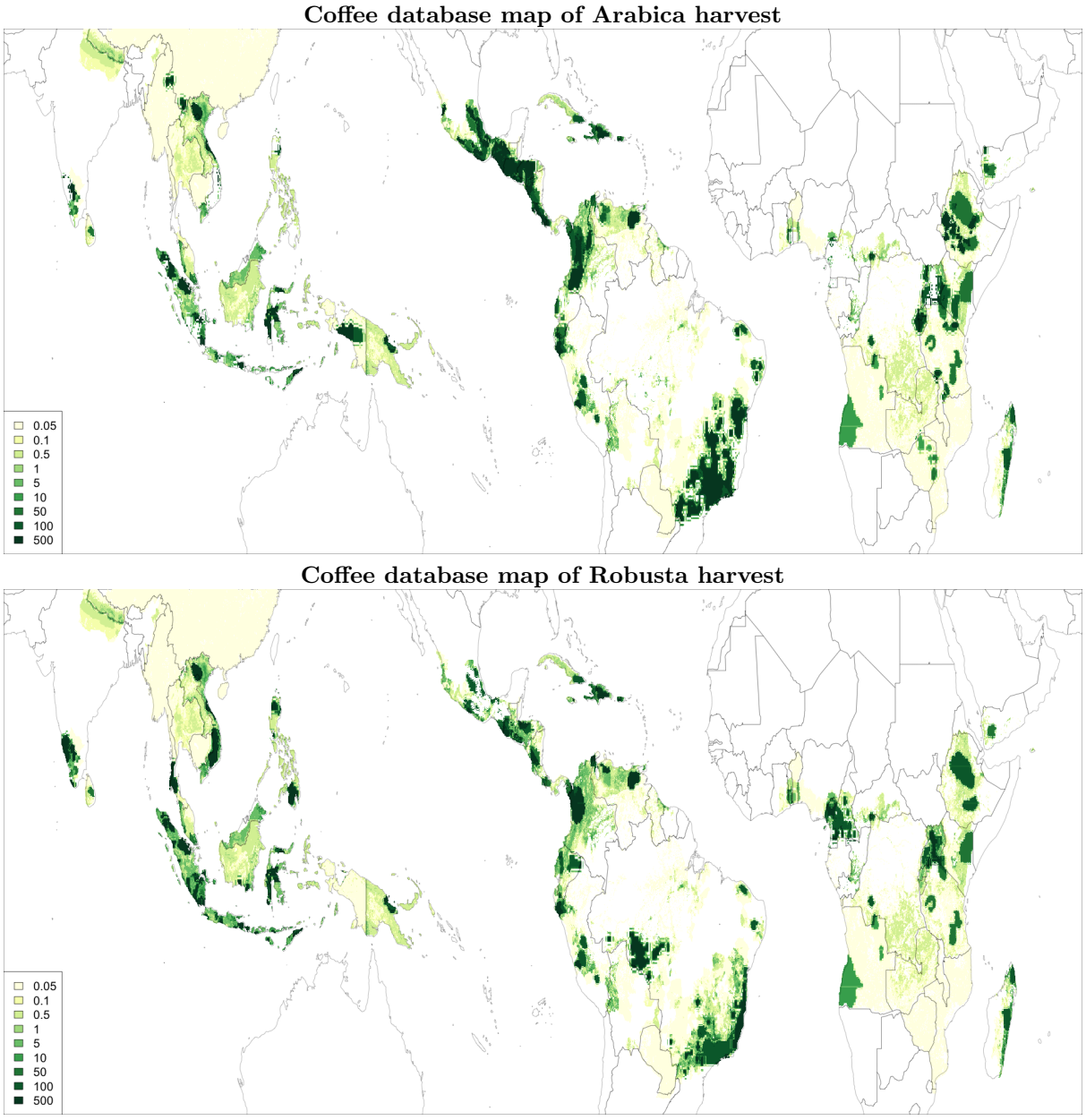


Figure 5: The spatial distribution of Arabica and Robusta harvests, as represented by spatial harvest maps. Harvest maps are combined across all months, and re-weighted so that the sum of grid cell values within a country is equal to the average harvested area in the most recent years of harvest.

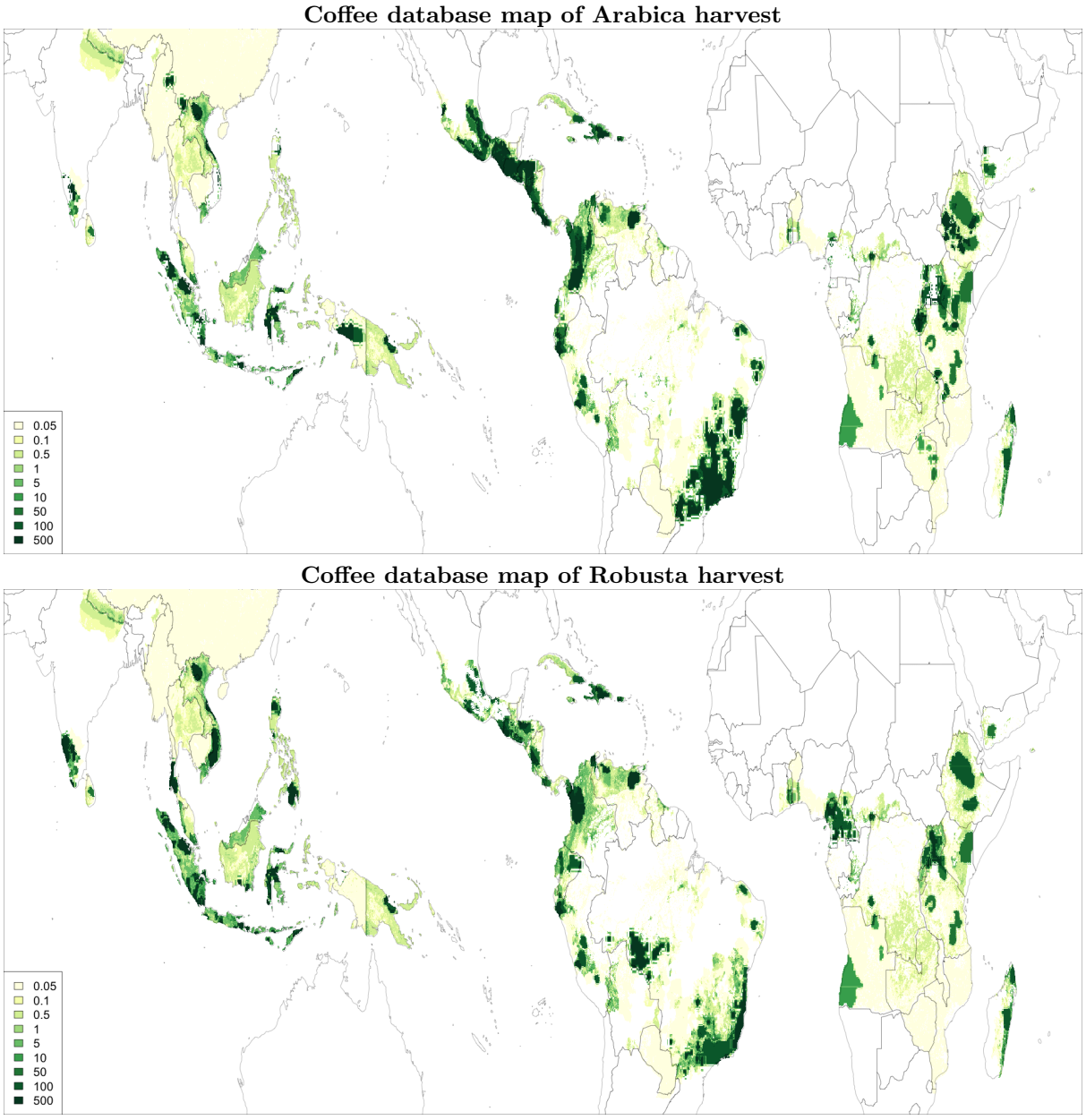


Figure 6: The spatial distribution of Arabica and Robusta harvests, as represented by spatial harvest maps. Harvest maps are combined across all months, and re-weighted so that the sum of grid cell values within a country is equal to the average harvested area in the most recent years of harvest.

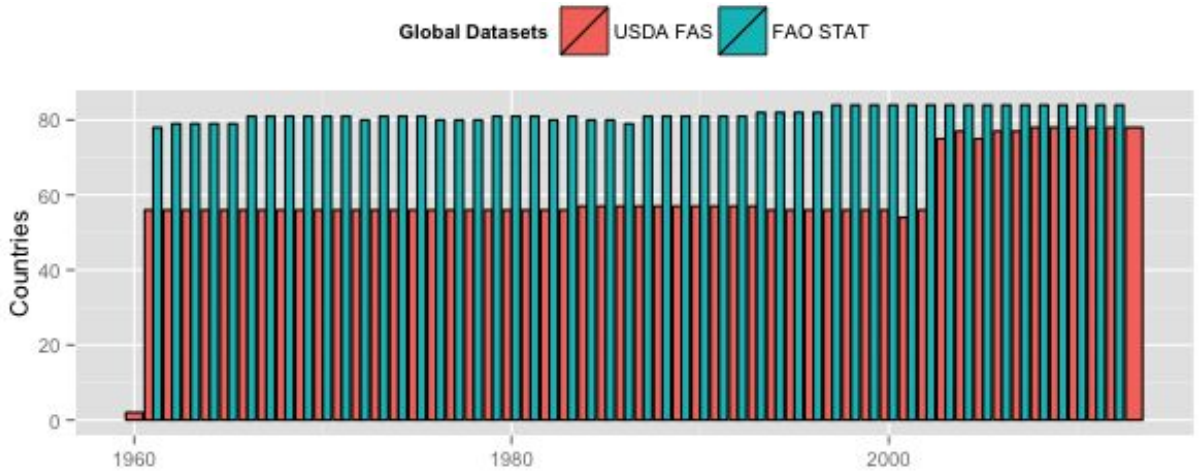


Figure 7: The number of countries with coffee production data available by year.

$USDA^R_- + USDA^L_+ = FAO$; that is, the FAO year consists of the 'right' (latter) portion of one market year and the 'left' (early) portion of the next one. Finally, we can use the difference to estimate α from

$$FAO_t - USDA_{t+} = \alpha(USDA_{t-} - USDA_{t+}) + \epsilon_t$$

We estimate this division for each country. In the case of Brazil, we find that 12% of production occurs between May and June, and 88% between July and September. A full table of these portions is shown in table 3. Where there are blanks, the two datasets could not be consistently combined.

Using these values, we can construct a monthly timeseries of production, as shown in figure 8.

Coffee production database The coffee database consists of paired production and growing region files. The database consists of both the final files and the code for generating standardized versions of input source files. The standardized versions have the same format as the merged database.

The coffee database is available in a sharable form, at `https://bitbucket.org/jrising/coffeedb/`. Request for access.

	Market Year	Previous Year	Following Year	Std. Err.
Brazil	Jul - Jun	0.12	0.88	0.05
Madagascar	Apr - Mar	0.50	0.50	0.25
Kenya	Oct - Sep	0.91	0.09	0.04
Guinea	Oct - Sep	0.37	0.63	0.33
Panama	Oct - Sep	0.68	0.32	0.28
Costa Rica	Oct - Sep	0.50	0.50	0.06
Ethiopia	Oct - Sep			
Rwanda	Apr - Mar	0.17	0.83	0.08
United Republic of Tanzania	Jul - Jun	0.67	0.33	0.11
Sri Lanka	Oct - Sep			
Peru	Apr - Mar	0.07	0.93	0.10
Lao People's Democratic Republic	Oct - Sep			
Bolivia (Plurinational State of)	Apr - Mar			
Cameroon	Oct - Sep	0.09	0.91	0.11
Côte d'Ivoire	Oct - Sep	0.89	0.11	0.07
Ecuador	Apr - Mar	0.41	0.59	0.25
Benin	Oct - Sep	0.60	0.40	0.20
Ghana	Oct - Sep	0.80	0.20	0.16
Cuba	Jul - Jun	0.46	0.54	0.16
El Salvador	Oct - Sep	0.40	0.60	0.05
Venezuela (Bolivarian Republic of)	Oct - Sep	0.57	0.43	0.15
Papua New Guinea	Apr - Mar	0.26	0.74	0.07
Malawi	Oct - Sep	0.10	0.90	0.16
Togo	Oct - Sep	0.44	0.56	0.16
Guatemala	Oct - Sep	0.42	0.58	0.17
Zimbabwe	Oct - Sep	0.19	0.81	0.10
Viet Nam	Oct - Sep	0.63	0.37	0.07
Dominican Republic	Jul - Jun	0.55	0.45	0.15
Nigeria	Oct - Sep	0.68	0.32	0.19
Liberia	Oct - Sep	0.56	0.44	0.14
Democratic Republic of the Congo	Oct - Sep	0.61	0.39	0.14
Paraguay	Oct - Sep	0.60	0.40	0.34
Trinidad and Tobago	Oct - Sep	0.77	0.23	0.10
Philippines	Jul - Jun			
Indonesia	Apr - Mar	0.23	0.77	0.29
Central African Republic	Oct - Sep	0.22	0.78	0.14
New Caledonia	Oct - Sep			
United States of America	Oct - Sep	0.19	0.81	0.65
Guyana	Oct - Sep			
Honduras	Oct - Sep	0.56	0.44	0.08
Yemen	Oct - Sep	0.36	0.64	0.86
Haiti	Jul - Jun	0.36	0.64	0.19
Thailand	Oct - Sep	0.77	0.23	0.07
Jamaica	Oct - Sep	0.69	0.31	0.98
Angola	Apr - Mar	0.62	0.38	0.11
Equatorial Guinea	Oct - Sep			
Mexico	Oct - Sep	0.62	0.38	0.22
India	Oct - Sep	0.98	0.02	0.02
Sierra Leone	Oct - Sep	0.98	0.02	0.60
Malaysia	Oct - Sep	0.58	0.42	0.33
Congo	Oct - Sep	0.28	0.72	0.26
Colombia	Oct - Sep	0.72	0.28	0.06
Burundi	Apr - Mar	0.06	0.94	0.09
Gabon	Oct - Sep	0.23	0.77	0.17
Uganda	Oct - Sep	0.83	0.17	0.09
Nicaragua	Oct - Sep	0.13	0.87	0.08
Zambia	Oct - Sep	0.93	0.07	0.35

Table 3: Portion of the production for each market year attributed to the previous calendar year and to the next one.

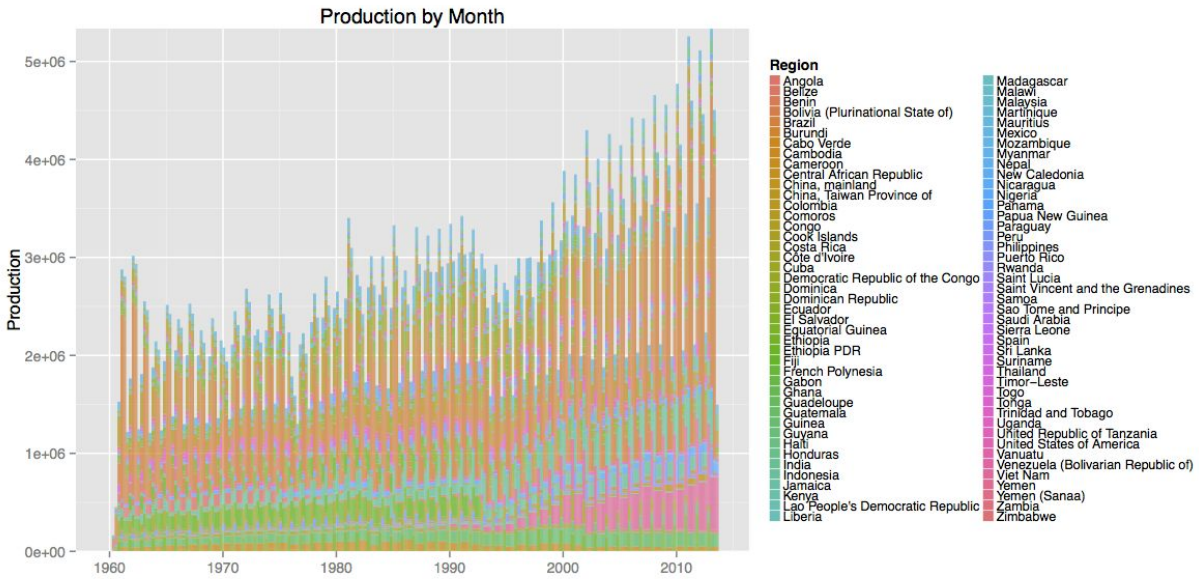
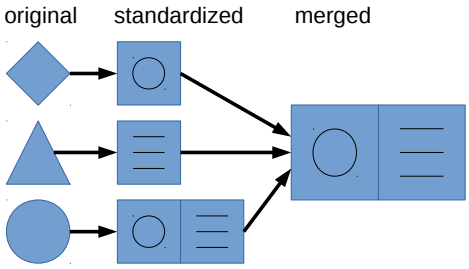


Figure 8: Production by month and country, inferred by the discrepancies between USDA FAS and FAO accounting systems.



A Standardized format

A.1 Growing Region Files

Growing regions are stored as raster (gridded) files describe "masks" of which regions are under harvest in a given month. They are at a resolution of 12 pixels per degree (a grid width of 5 minutes), and cover the entire area from 180°W to 180°E longitude, and 30°S to 30°N latitude.

The grids are stored as NetCDF files. Each NetCDF file contains a "harvest" variable and a "confidence" variable. The harvest variable specifies areas under harvest in a given month, and has dimensions Longitude x Latitude x 12, with a separate mask for each month. Values may range between 0 and 1, based on how much evidence there is of harvest there. The confidence mask describes the level of confidence in the information, also from 0 to 1.

Note that neither the "harvest" nor "confidence" variables describe the portion of a given grid cell under harvest. Instead, both relate to the evidence that the grid cell contains areas under harvest. The difference between the harvest value and confidence value is expanded upon in the Merging Growing Region Files section.

A.2 Production Files

Production data consists of a `.csv` file that specifies the production, planted area, harvested area, and yields (as data is available for each) in a given year and a given region. Where the data describes sub-country regions, an additional region definition file (`*-regions.csv`) and a shapefile database (collections of a `.shp`, `.shx`, and `.dbf` file) describe an association between the production records and growing regions. Each polygon in the shapefile database identifies a region for which production data is available in one or more years, and region definition files specify which region is described in each record.

The production file has the following column header:

`year,region,variety,produced,prod-se,harvested,harv-se,planted,plant-se,yield,yield-se`

`year` is the year being described. Not all years need to be represented for a region. `variety` is Arabica, Robusta, or combined. `region` is the region identifier in the associated region definitions file. This may change across years. `produced` is the calendar year production, measured in metric tonnes. `prod-se` is the standard error of the production estimate. It may be `NA` if the error estimate is available, but this will cause any other estimate to be chosen over it if one is available. `harvested` is the harvested area in hectares, and `harv-se` is its standard error. This may be `NA`. `planted` is the planted area in hectares, and `plan-se` is its standard error. This may be `NA`. `yield` is the yield in terms of MT per hectare, and `yield-se` is its standard error. The yield is computed as production divided by planted area. This may be `NA`.

The region definitions file has the following column header:

`region,PID,weight`

`region` is a region identifier, unique across the entire database. If there are multiple rows with the same region identifier, all of these PIDs will be combined in the region. `PID` is a polygon IDs in the associated growing region file. The same `PID` may occur in multiple regions, since different regions may be used to describe different years. `weight` is a measure of the accuracy of the production region definitions. In general, `weight` is calculated as *(mean planted area) / (total polygon area)*, and is between 0 (no confidence) and 1 (full confidence).

B Generating merged database files

B.1 Merging growing region files

Region definition files are merged according to the weight of evidence of harvest in each month. At every point, the new weight in the combined region definitions is,

$$w(x,y) = \sum_i \frac{w_i(x,y)c_i(x,y)}{c_i(x,y)}$$
$$c(x,y) = \sum_i c_i(x,y)$$

This formulation allows confidence to increase where multiple data sources are available, and causes contradictory maps (for example, one that says that coffee is grown in a region and one that says it is not) to result in averaged values.

	Arabica 1	Arabica 0	Combined 1	Combined 0
Arabica 1, Robusta 0	1, 0	0.5, 0	1, 0	0.5, 0
Arabica 0, Robusta 0	0.5, 0	0, 0	0.5, 0.5	0, 0

Additional logic is used where maps that describe Arabica and Robusta growth separately are combined with those that lump them together.

B.2 Merging production files

Production files are merged using a Bayesian approach, with a uniform prior. Each estimate (a given year-region value for production, harvested area, or planted area) is translated into a distribution, $p(y_i)$. Then the merged estimate of production is,

$$p(y) = \prod_i p(y_i)$$

This allows the database to account for uncertainty in the estimates, as well as allowing corroborating records to decrease the amount of uncertainty. Additional logic used where time series that split out Arabica and Robusta growth (such as the USDA Foreign Agricultural Service) are combined with ones that lump them together (such as FAO).

Combining estimates

1. Associate uncertainty with each observation ($k\sqrt{v_{it}}$).
2. Case 1: Same or Combined + (Arabica or Robusta):

$$\mu* = \frac{\frac{\mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}$$
$$\sigma*^2 = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}$$

3. Case 2: Combined + Arabica + Robusta:

$$\max_{a,r} \mathcal{N}(a|\mu_a,\sigma_a)\mathcal{N}(r|\mu_r,\sigma_r)\mathcal{N}(a+r|\mu_c,\sigma_c)$$

σ_a* and σ_r* from Inverse Hessian.

C Production maps

We performed a geospatial matching between diagrams in the gray literature and countries maps.

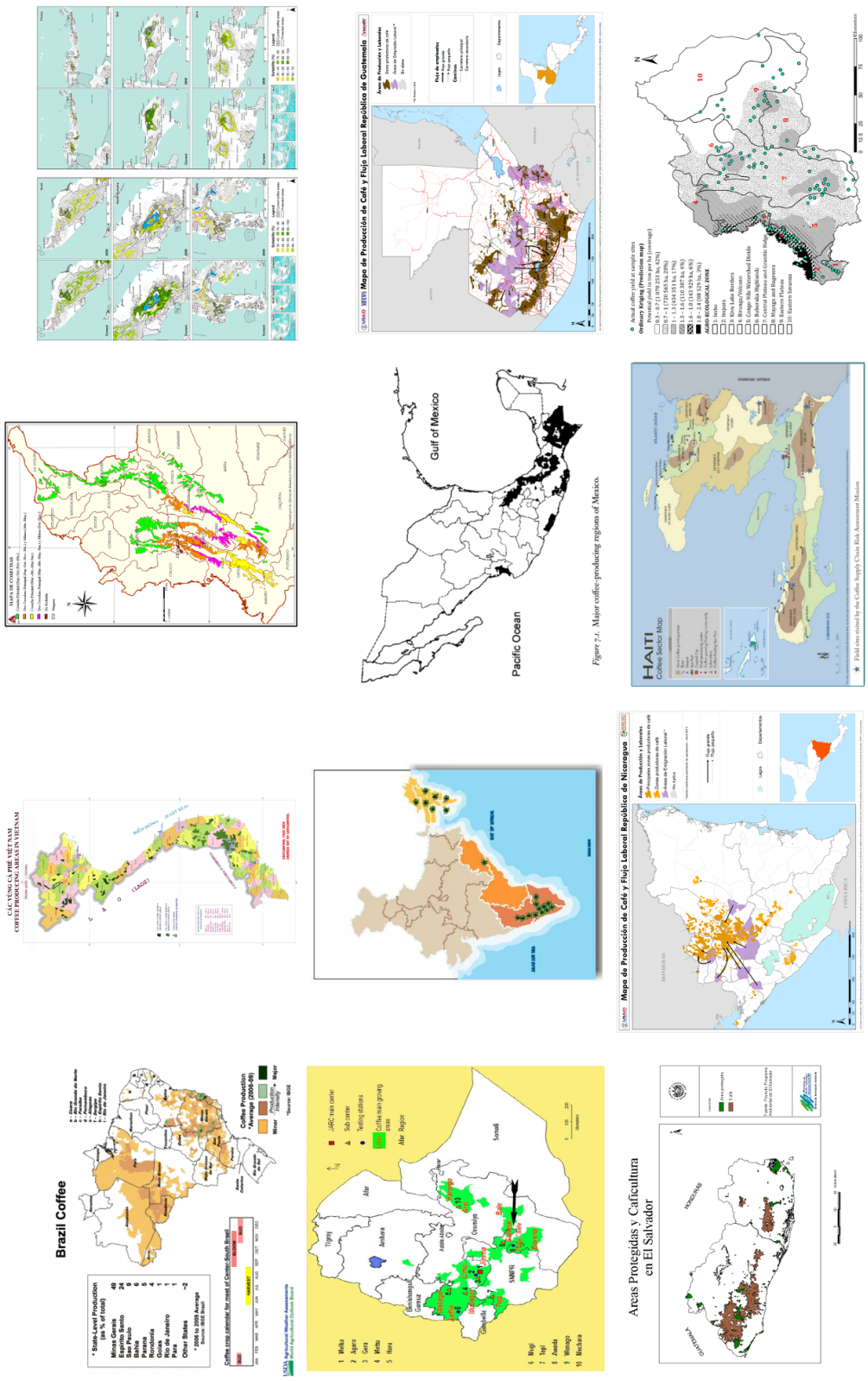


Figure 7.1. Major coffee-producing regions of Mexico.

Figure 9: Coffee growing region maps for Brazil, Vietnam, Colombia, Indonesia, Ethiopia, India, Mexico, Guatemala, El Salvador, Nicaragua, Haiti, and Rwanda, collected from various sources (see main text for the table).

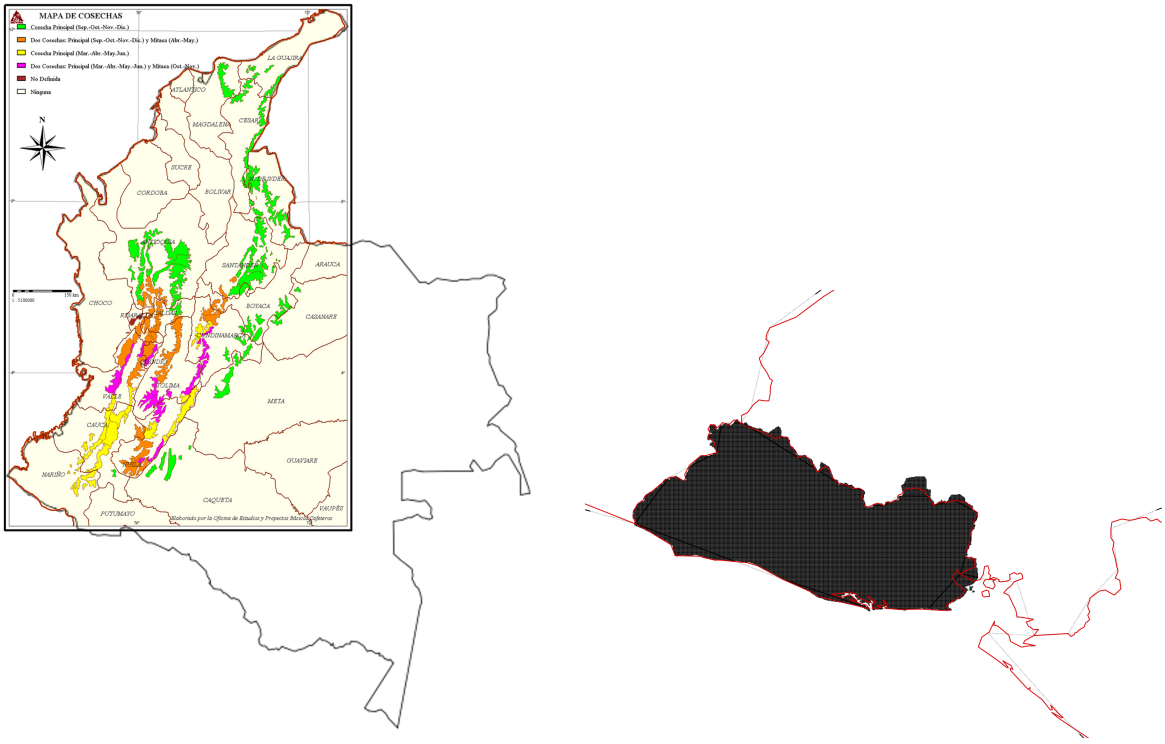


Figure 10: Two examples of the geospatial matching process, using hand correspondences for Colombia (left) and country-wide shape matching for El Salvador (right).

References

Bunn, C., Läderach, P., Rivera, O. O., and Kirschke, D. (2015). A bitter cup: climate change profile of global production of arabica and robusta coffee. *Climatic Change*, 129(1-2):89–101.

Gonzalez, R. J. (2010). *Zapotec science: Farming and food in the Northern Sierra of Oaxaca*. University of Texas Press.

Maxey, M. (2015). Arabica coffee from yemen: Hope in a time of turmoil. Available at `https://www.linkedin.com/pulse/arabica-coffee-from-yemen-hope-time-turmoil-michael-maxey`.

Monfreda, C., Ramankutty, N., and Foley, J. A. (2008). Farming the planet: 2. geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000. *Global biogeochemical cycles*, 22(1).

Sweet Maria (2015). World coffee production timetable. Available at `http://sweetmarias.com/coffee.prod.timetable.php`.

Thurston, R. W., Morris, J., and Steiman, S. (2013). *Coffee: A Comprehensive Guide to the Bean, the Beverage, and the Industry*. Rowman & Littlefield Publishers.

Acknowledgments

We would like to thank **Walter Baethgen** at the International Research Institute for Climate and Society for his thoughtful reviews of the work here and all of his comments and suggestions.

Many thanks also to:

Andrea Illy, illycaffè S.p.A., Chairman/CEO

Mario Cerutti, Lavazza S.p.A., Corporate Relations

Belay Begashaw, Columbia Global Centers, Director

Mauricio Galindo, International Coffee Organization, Head of Operations

Alexandra Tunistra, Root Capital, Advisory Services

Amir Jina, University of Chicago, Economics

Joann de Zegher, Stanford University, Environment and Resources

Marion Dumas, Columbia University, School of International and Public Affairs