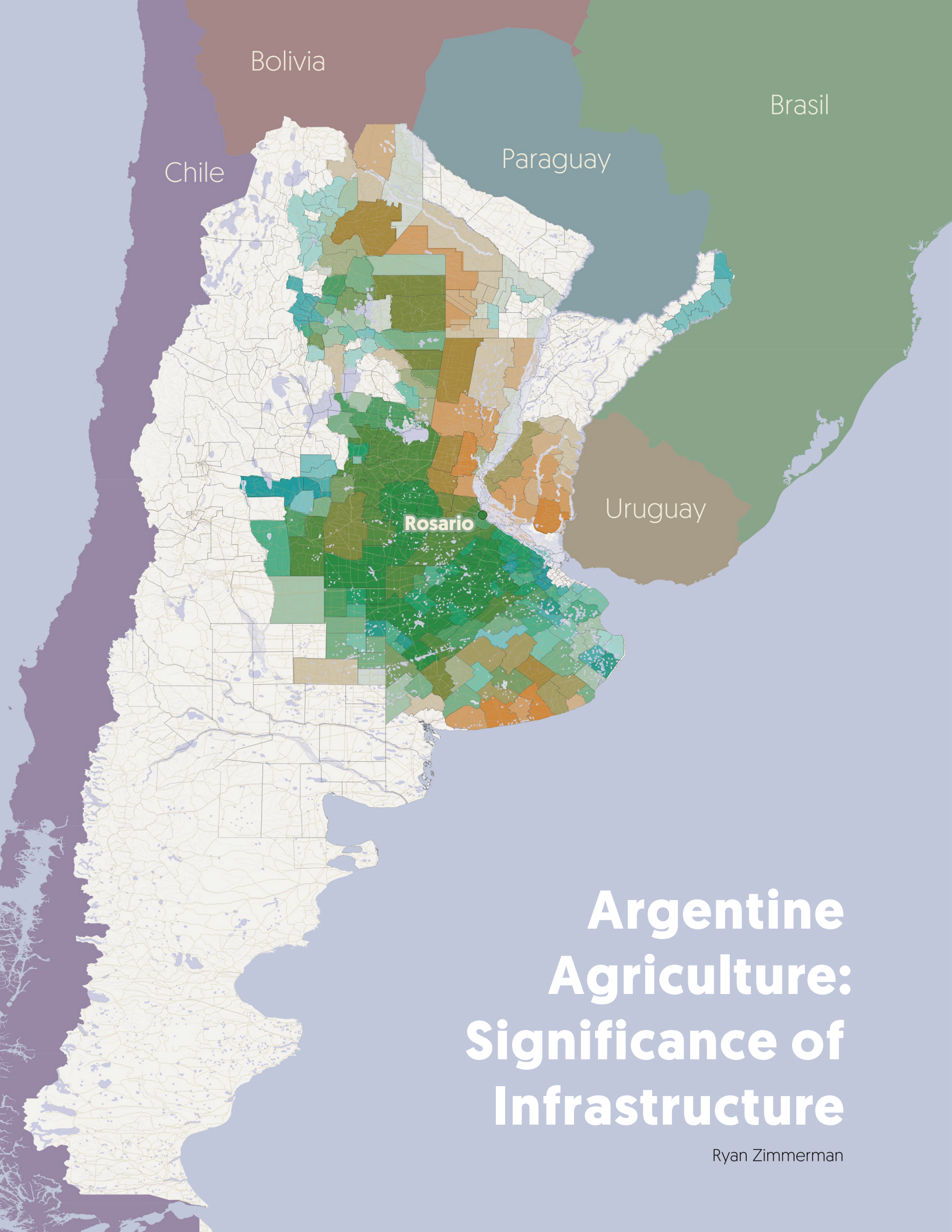


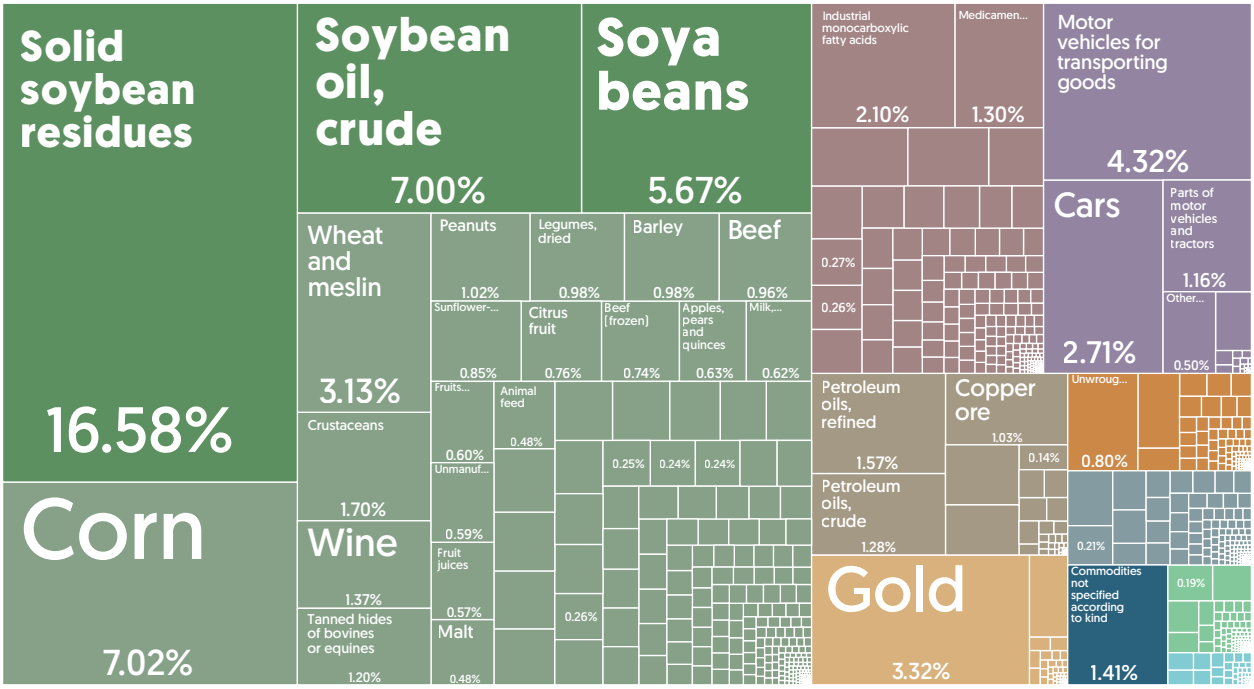
Argentine Agriculture: Significance of Infrastructure

Ryan Zimmerman

Why Soy?

Agricultural products form 65% of Argentina's exports, and a full 30% of that agricultural product is raw or refined soy, making it the single most important export crop in Argentina.[1] In addition, a majority of the soy production is exported because it is not a traditional food-crop—most soy planted is planted with the world market in mind. Furthermore, the vast majority of soy crop harvested is processed before export in just a few locations, the formost being the city of Rosario. This simplicity makes it a fascinating crop to model, because correlations with primary variables are readily apparent.

Argentinian Gross Exports by Type:



[1] **Source:** *What did Argentina Export in 2016?* Atlas of Economic Complexity, Center for International Development at Harvard University

Key Factor: Yield

The first factor I wanted analize in developing a model for soy planting is the yield rate: the amount of soy it is possible to produce in a given plot of land. This is only one of the factors farmers face when deciding which crop to plant, but it is a critical one because a farmers profitability is a function of their yield quantity for a given crop, multiplied by the price of the crop, after subtracting their total costs of planting, tending, harvesting, and selling the crop.

Fortunately, both yield rates and production quantities are published per-district by the Argentine government as part of their Datos Abiertos program which started in 2012 to improve civic engagement and government transparency. These data make it possible to see discrepancies between where farmers are likely to get the largest harvest quantity for a given amount of land and how much soy is actually produced in each district. I expected to see a strong correlation between districts with the highest yield rates and districts with the highest production volume.

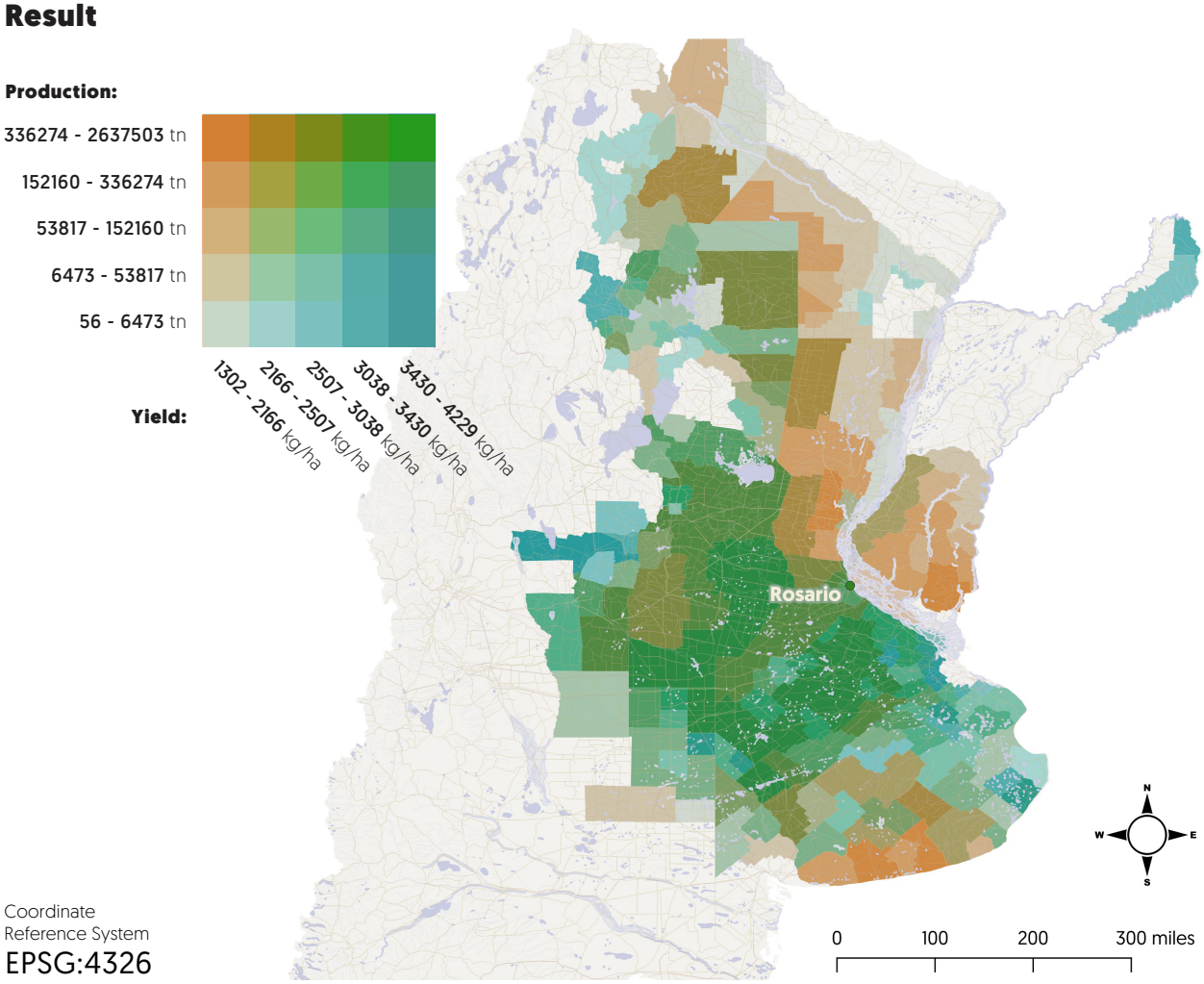
Bivariate Choropleth Map

To identify regions where soy yield and production are more and less correlated, I built a bivariate choropleth map with production quantity on one choropleth (the orange axis) and soy yield rates on the other (the blue axis). Where the colors overlap, they are combined by selecting the minimum values of each color. In other words:

$$rgb = [\min(r_1,r_2),\ \min(g_1,g_2),\ \min(b_1,b_2)]$$

Where rgb_1 and rgb_2 are the positions on the orange and blue color ramps respectively. The result is a third color, green, which appears wherever there is high saturation of both blue and orange: places where there is both high production and yield.

Result



This map makes it easy to identify two types of areas where further study would be especially interesting: orange tinted areas where farmers are managing to profit despite lower yields, and blue-tinted areas where soy is not grown despite high potential yields. The strategic significance of Rosario is immediately apparent, because it is located on a navigable section of the Río Paraná as close as possible to the highest yield areas. Likely because Argentina's soy crushing capacity is located in rosario, most of the blue-tinted areas are far from Rosario, implying that transportation costs play a large role in farmer's profits. Likewise, most of the orange low-yield, high production areas are near Rosario,

which suggests that their lower transport costs might be helping them stay profitable in the face of competition with higher yields. However, there are regions on the north and south of the map which have high production despite being far from Rosario, so there are likely further factors at play in those regions.

Method

I used the open-source software QGIS to read and manipulate the .shp files provided by the Ministerio Agroindustria Visor IDE tool, which are spreadsheet-like files containing rows of features (in this case districts) which each contain attributes (such as yield) and a polygon which represents the area the feature covers. These files connect numerical data to geospatial areas, making it possible to visualize the data and perform calculations based on area and distance. To make the yield and production data easier to understand, I built a basemap using province and department boundaries from the GADM database, and retrieved Argentine waterways, water areas, roads, and railways databases from the 1992 Digital Chart of the World through the DIVA-GIS site.

Next Steps

The choropleth map clearly illustrates how insufficient yield rates alone are for modeling the planting decisions of farmers, but the visualization is a useful tool for designing a future model which would account for transport distance of produced goods and opportunity costs of planting other crops.

Sources:

Visor IDE, Ministerio de AgroIndustrial de Argentina, ide.agroindustria.gob.ar/visor
Global Administrative Areas, Robert Hijmans, www.gadm.org
Digital Chart of the World, retrieved from www.diva-gis.org/gdata
What did Argentina Export in 2016? Atlas of Economic Complexity, Center for International Development at Harvard University

Essay 2 | Argentinian Soybeans: **Processing Infrastructure**

Soy: Path to Market

Once harvested, Soybeans require processing before they can be used for food, animal feed, or biodiesel. Therefore, farmers must consider the costs of transporting their harvest to wherever it will be sold or processed. Over 75% of soybeans produced in Argentina are crushed for either meal or oil, in part because of differential export taxes: soybeans are taxed more heavily upon export than meal or oil. Most of the soybean meal produced is exported, but there is some domestic consumption as livestock feed, especially by the pork industry (Mergen, Sandoval, 2017). The soybeans that aren't crushed are largely sold on a grain market, primarily the Bolsa De Commercio Rosario.

Key Factor: Crushing Infrastructure

Because crushing represents such massive demand for soy—and is vast majority of domestic consumption—the locations and sizes of soy crushing plants will play a large role in farmers' decisionmaking process to plant soy; being near a high-capacity crushing plant makes it easy to transport and sell the crop. Crushing plants generally fall in to one of two types: huge export-focused plants located along the river Parana, and smaller, newer plants located inland. The smaller inland export plants focus on soybean meal production for domestic livestock feed market, while the large-scale plants operate at a much larger scale and load most of that product immediately onto ships for export. The smaller plants can be located in the middle of high-yield farmland, whereas the larger plants are concentrated in Rosario to minimize transportation costs to the international market (Mergen et al, 2017).

Renova Timbues
source: www.renova.com.ar

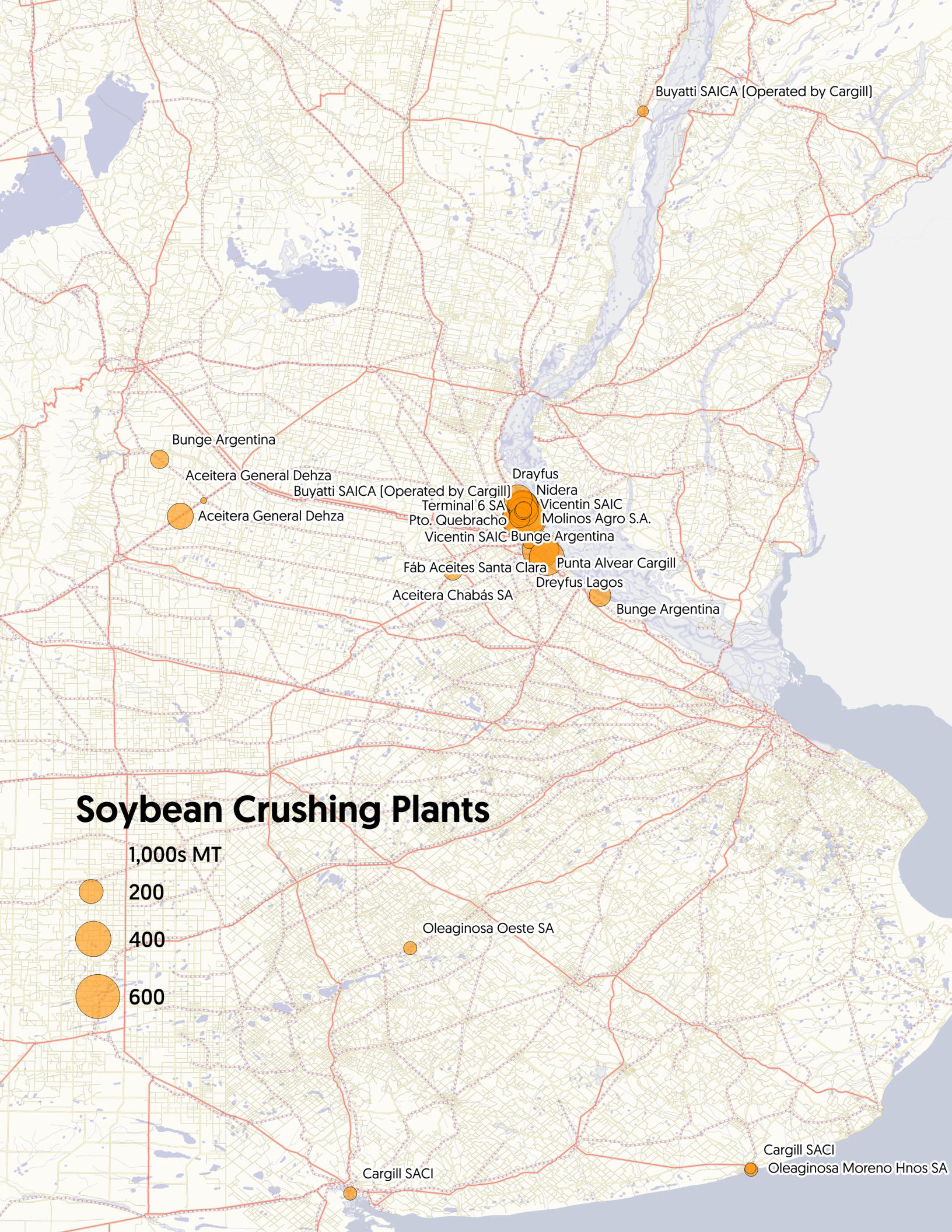
Crushing Plant Locations

When examining soybean crushing plants, especially the larger ones such as the Renova Timbues plant above, the strategic significance of the city of Rosario is readily apparent. The river Parana is navigable and has plenty of ports all the way up to Santa Fe, but Rosario is significant because it is nearer to the high-yield farmland to the west. For that reason, about 80% of soybean crushing capacity is located in Rosario.

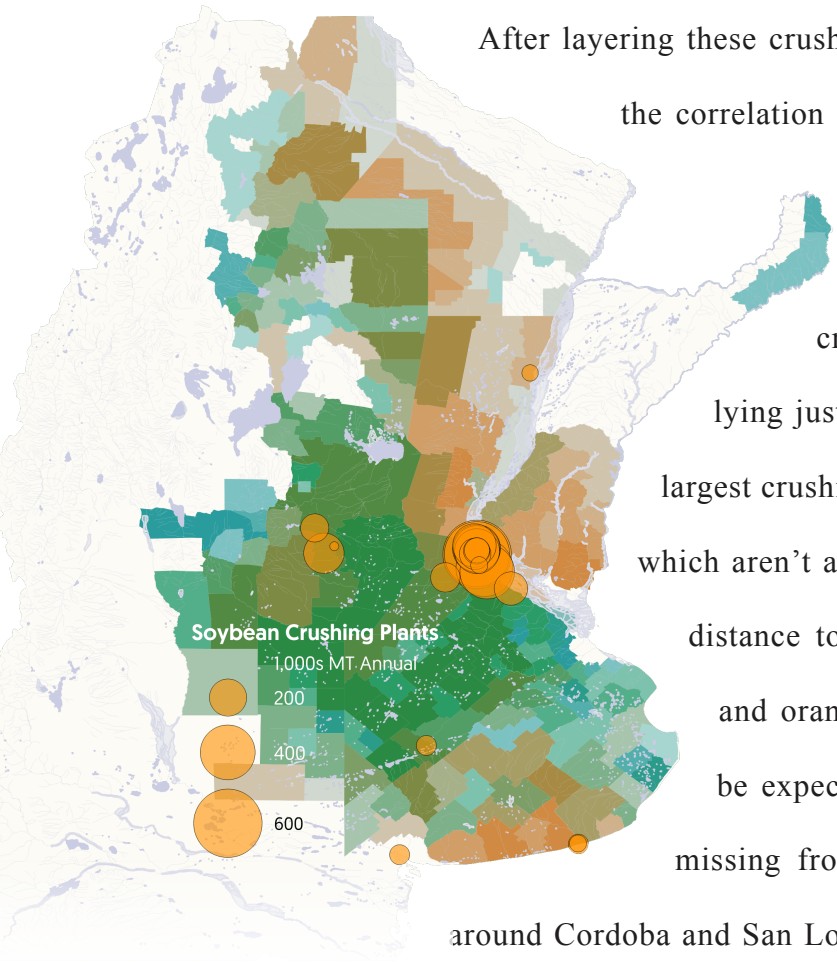
To find the locations and capacities of the crushing plants, I used the operations manual published by the Bolsa De Commercio De Rosario. This source shows the plant and their capacities as it was in 2008, but it is a huge limitation and flaw in this preliminary research because more plants have been built (such as the massive Renova plant above; construction started in 2010 operations began in 2013) and the yield and production data I am using are much more recent. After copying each crushing plant out of the book, I used google satelite imagery to find each plant's exact location and create a geospatial point file to use for later calculations. The next two pages show that research.

Argentina Crushing Plants and Capacities

Plant	City	24 Hour Capacity	Annual Cap.
Pto. Quebracho	Puerto General San Martín	9000	2970000
Cargill SACI	Ingeniero White	2000	660000
Cargill SACI	Necochea	2000	660000
Punta Alvear Cargill	Villa Gobernador Gàlvez	13000	4290000
Terminal 6 SA	Puerto General San Martín	9500	3135000
Bunge Argentina	Puerto General San Martín	8500	2805000
Bunge Argentina	San Jeronimo Sur	2200	726000
Bunge Argentina	Tancacha	3700	1221000
Bunge Argentina	Ramallo	5000	1650000
Vicentin SAIC	San Lorenzo	16500	5445000
Vicentin SAIC	Ricardone	4500	1485000
Terminal 6 SA	Puerto General San Martín	9500	3135000
Aceitera Chabàs SA	Chabás	4000	1320000
Aceitera General Dehza	General Deheza	7000	2310000
Aceitera General Dehza	D. Vélez Sársfield	500	165000
Dreyfus Lagos	General Lagos	12000	3960000
Drayfus	Timbúes	8000	2640000
Fàb Aceites Santa Clara	Rosario	1500	495000
Molinos Agro S.A.	San Lorenzo	18000	5940000
Oleaginosa Oeste SA	General Villegas	2000	660000
Oleaginosa Oeste SA	Daireaux	2000	660000
Oleaginosa Moreno Hnos SA	Necochea	1500	495000
Aceites GRAINER	Paraná	600	198000
Nidera	Puerto General San Martín	3000	990000
Nidera	Saforcada (Junín)	2500	825000
Buyatti SAICA (Op. Cargill)	Puerto General San Martín	3300	1089000
Buyatti SAICA (Op. Cargill)	Reconquista	1400	462000


Buyatti SAICA (Operated by Cargill)
Bunge Argentina
Aceitera General Dehza
Buyatti SAICA (Operated by Cargill)
Aceitera General Dehza
Drayfus
Nidera
Terminal 6 SA
Vicentin SAIC
Pto. Quebracho
Molinos Agro S.A.
Vicentin SAIC
Bunge Argentina
Fáb Aceites Santa Clara
Punta Alvear Cargill
Dreyfus Lagos
Aceitera Chabás SA
Bunge Argentina
Soybean Crushing Plants
1,000s MT
200
400
600
Oleaginosa Oeste SA
Cargill SACI
Cargill SACI
Oleaginosa Moreno Hnos SA

Result



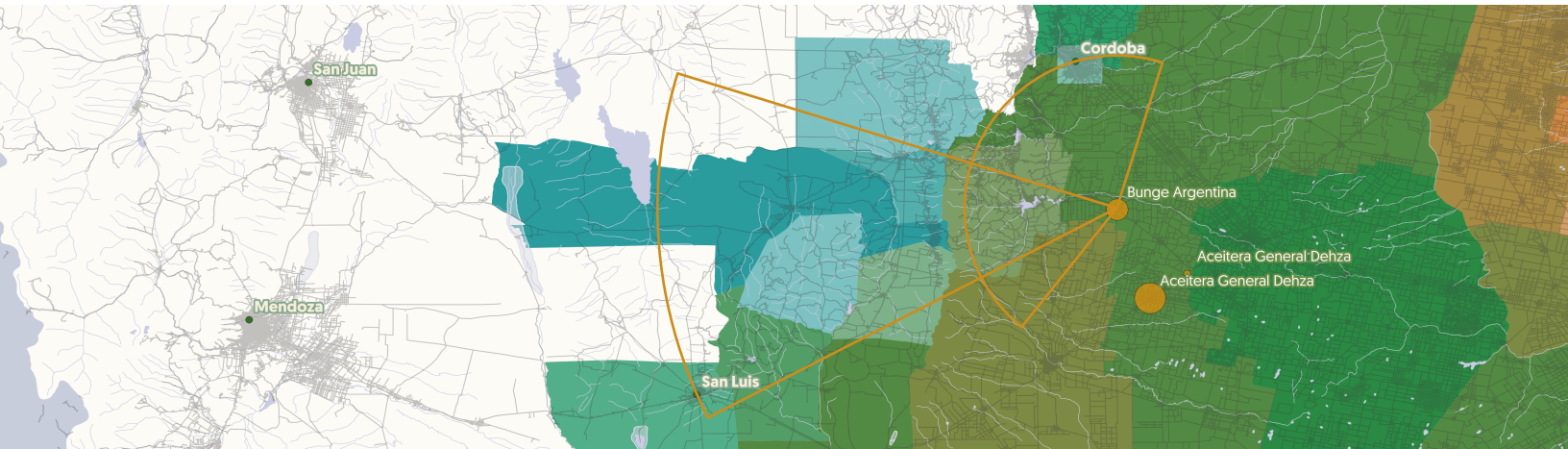
After layering these crushing plants on the bivariate choropleth map, the correlation is visible: the orange areas, where farmers are producing soy despite relatively lower yield, are almost universally near soybean crushing plants, with the largest orange area lying just to the east of Rosario, which contains the largest crushing capacity. However, there are a few areas which aren't adequately explained by simple straight-line distance to crushing capacity. There are many green and orange districts in the north where blue would be expected, potentially indicating a crushing plant missing from my data. The other unexpected area is around Cordoba and San Lois; constant radii from the nearest crushing plant drawn through those cities pases through districts where production is both higher and lower than yield would predict, indicating that there is some other variable to consider.

Sources:

Visor IDE, Ministerio de AgroIndustrial de Argentina, ide.agroindustria.gob.ar/visor

Global Administrative Areas, Robert Hijmans, www.gadm.org

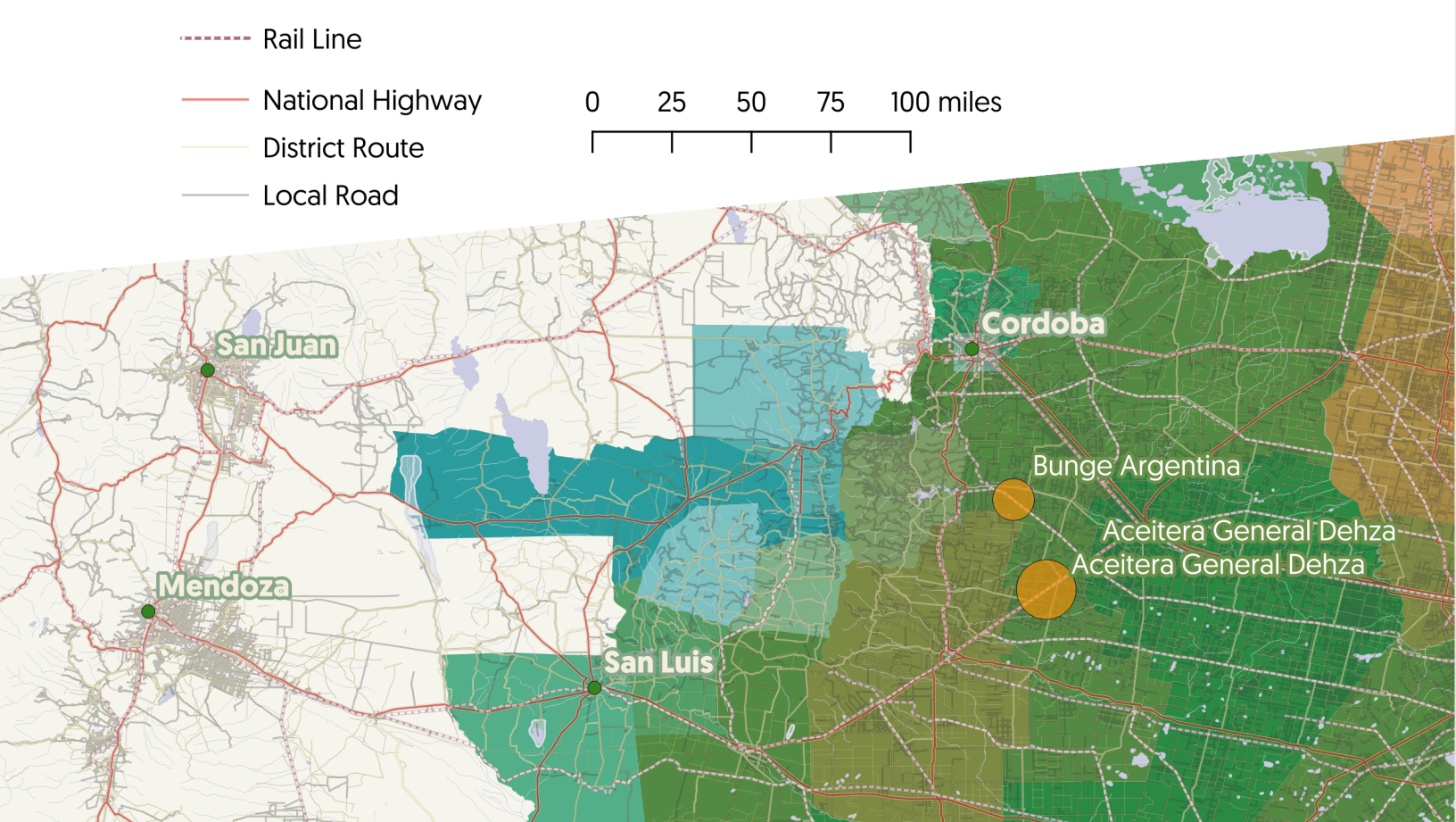
Mergen, David, and Lazaro Sandoval. "Oilseeds and Products Annual 2017/18 Forecast: Limited area growth for Soybeans, Sunflower, and Peanuts." GAIN Report, 26 Apr. 2017. Gain.fas.usda.gov.



Transportation Network Effects

The results from Essay 2 clearly show that across much of the map, especially the southern half, a strong connection is visible between areas where production is higher than yield would predict and soybean crushing plant locations. However, there are some regions where the connection is not readily apparent, most notably the districts between Cordoba and San Luis, where there are blue districts at the same straight-line distance from the crushing facility as neighboring green and orange districts.

However, after highlighting the district and national highway system, it becomes clear that from a practical perspective—where grain is shipped by trucks and trains instead of airplanes or helicopters—the midpoint between the two towns is almost twice as far by

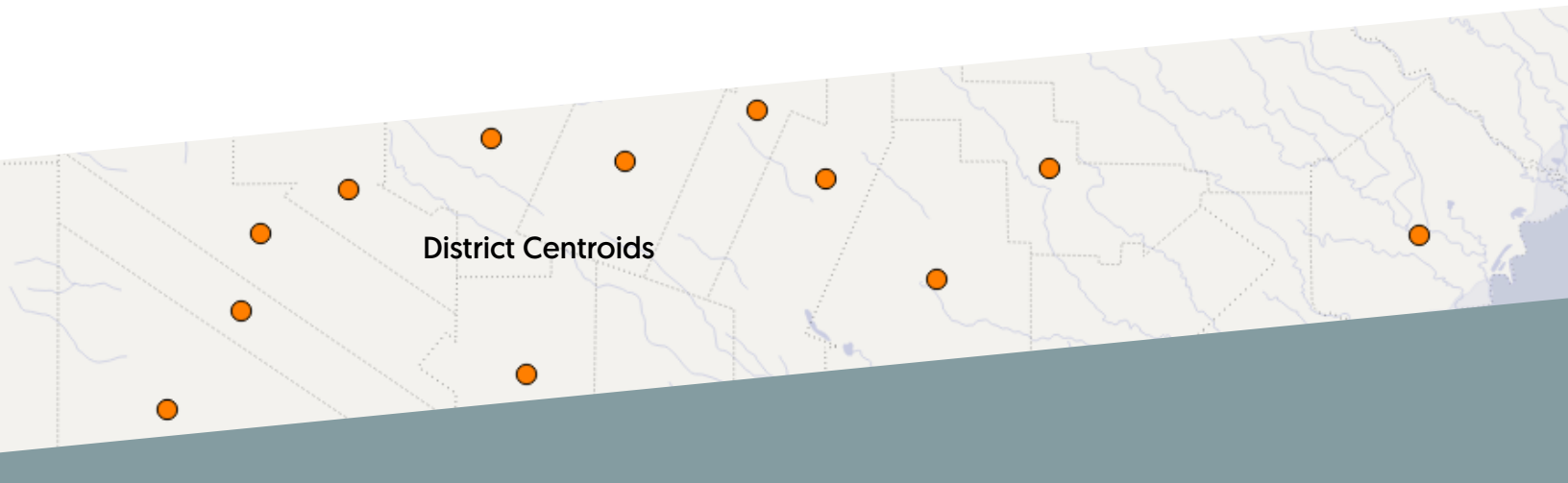


car or truck than in a straight line because you have to first drive to one of the towns before turning on to a highway headed towards the soybean crushing plant. This means that further analysis needs to be based on distances according to the road system, which requires network analysis to determine optimal driving routes from each soy-producing district to the soy crushing plants.

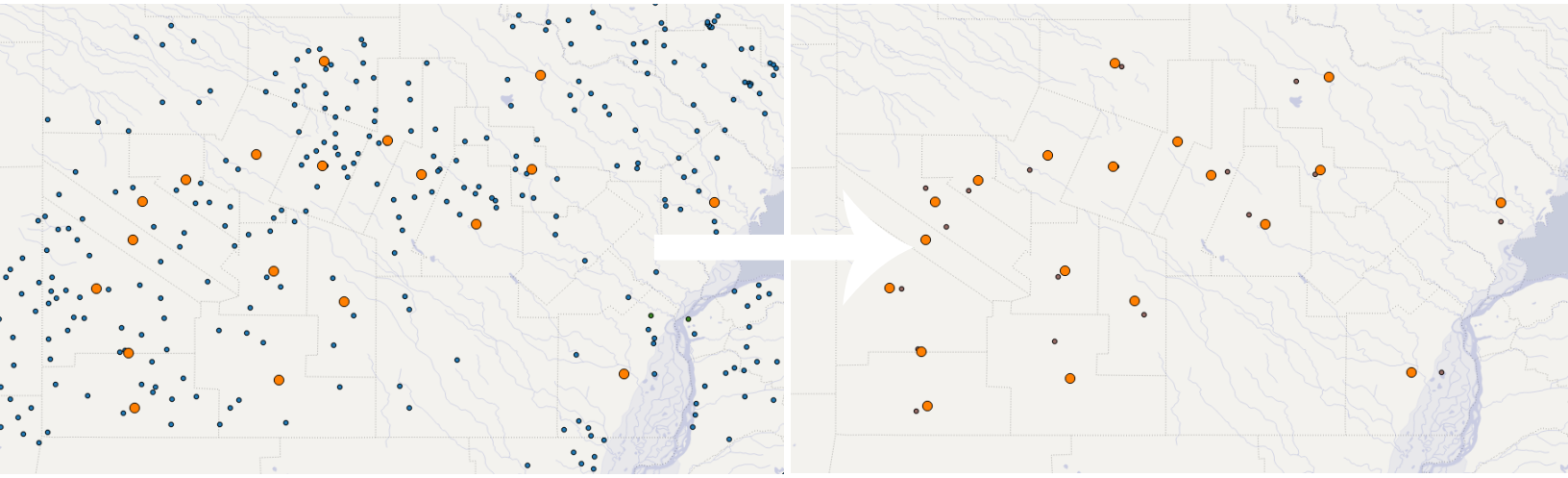
Method

From a technical standpoint, calculating every route by road from each soy-producing district to each crushing plant is non-trivial. The soy yield and production data provided by the Ministerio de AgroIndustrial de Argentina covers 285 districts, and my table of soy crushers based on the Bolsa De Commercio handbook contains 23 facilities, which means the network analysis will need to produce 6,555 routes to connect each start and end point.

The first challenge was to identify 285 start points which would be roughly representative of an average farmer in the district delivering soy (or paying for delivery). To do this, I first make the simplyifying assumption that each district contains an even distribution of farms, and that transportation cost from any point in the district to any other point follows the same linear function—in other words, an even distribution of farms connected by an even distribution of roads all of the same quality. If these assumptions are acceptable, then the centroid of the district should represent the average travel time from every point in the district to an arbitrary point outside it, in this case a soybean crushing plant.



However, these centroids are based only on the shape of the districts, which means that they may not be close enough to roads to be valid start points for network analysis. To increase the likelihood that a network analysis tool would be able to find a road, I ran a near-est-neighbor analysis on the “Centro Poblados” data provided by the Instituto Geográfico Nacional to identify the nearest populated place to the centroid of each district.



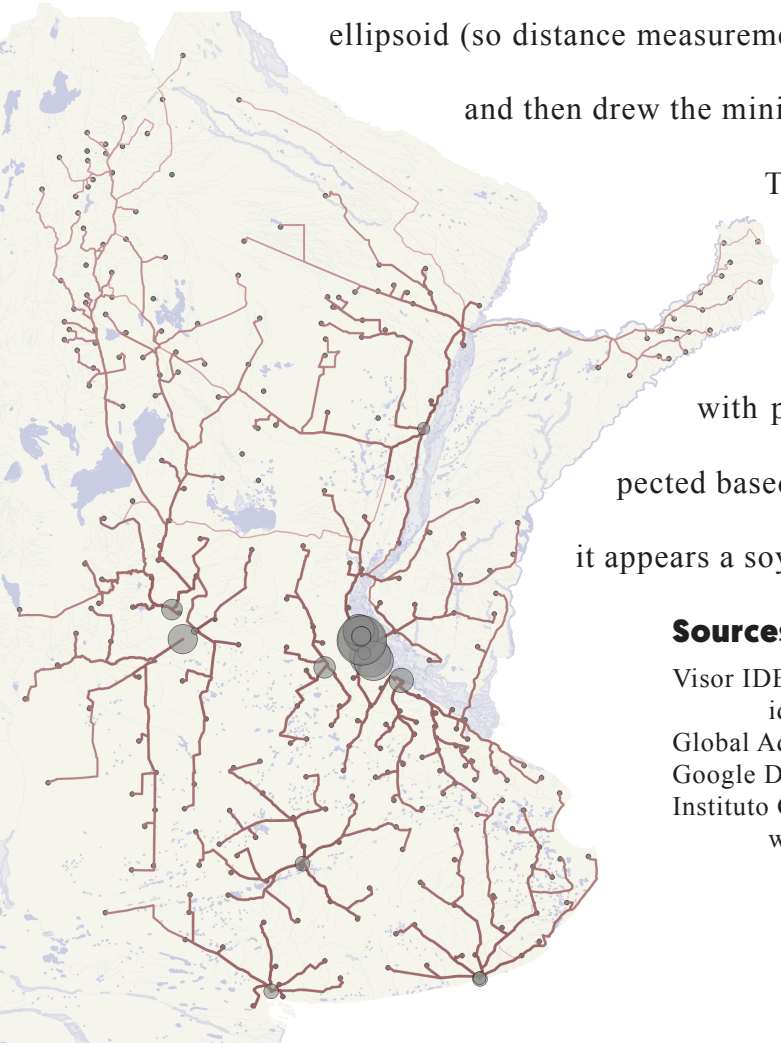
Unfortunately, the “Centro Poblados” data provided does not have detail beyond name and whether a feature is a province captial, department capitol, or ‘other.’ This makes it impossible to set criteria for which nearby population centers to use beyond whether or not the Instituto Geográfico Nacional decided to include them in the file.

The first system I attempted to use for road network analysis was Open Route Service, through a QGIS plugin that was able to read my data tables straight from QGIS. However, at a limit of 2,500 requests per day, it would take several days to get the data I needed. In order to escape the request limits of commercial routing services, I turned to an open-source program called Graphhopper. Becuse Graphhopper is open source, I was able to host the web server on my own computer, which has the added benefit of not requiring an internet

connection to make API calls once the road data had been downloaded from Open Street Map. Unfortunately, after getting Graphhopper running, I found that the Open Street Map data it uses does not have sufficient coverage for rural Argentina (in hindsight, this would have been a problem for Open Route Service as well) so even after picking 'populated' points many of them were too far from a road in the OSM database to be a valid start point.

The search for more complete road data lead me back to proprietary, commercial solutions. Google maps' coverage is excelent, even in rural areas, and their 2,500 API calls per day limit can be lifted for a small fee. I wrote a python script to produce all the API calls, and then used a separate script by Github user signed0 to decode the polyline result to a list of coordinate pairs, which I saved to a csv file for further processing.

Finally, I wrote one last script which analyzed every line in the csv file, measured the distance between each point in each path using the vincenty formula with the WGS-84 ellipsoid (so distance measurements would be the same as those done in QGIS) and then drew the minimum distance path on the the QGIS layer.

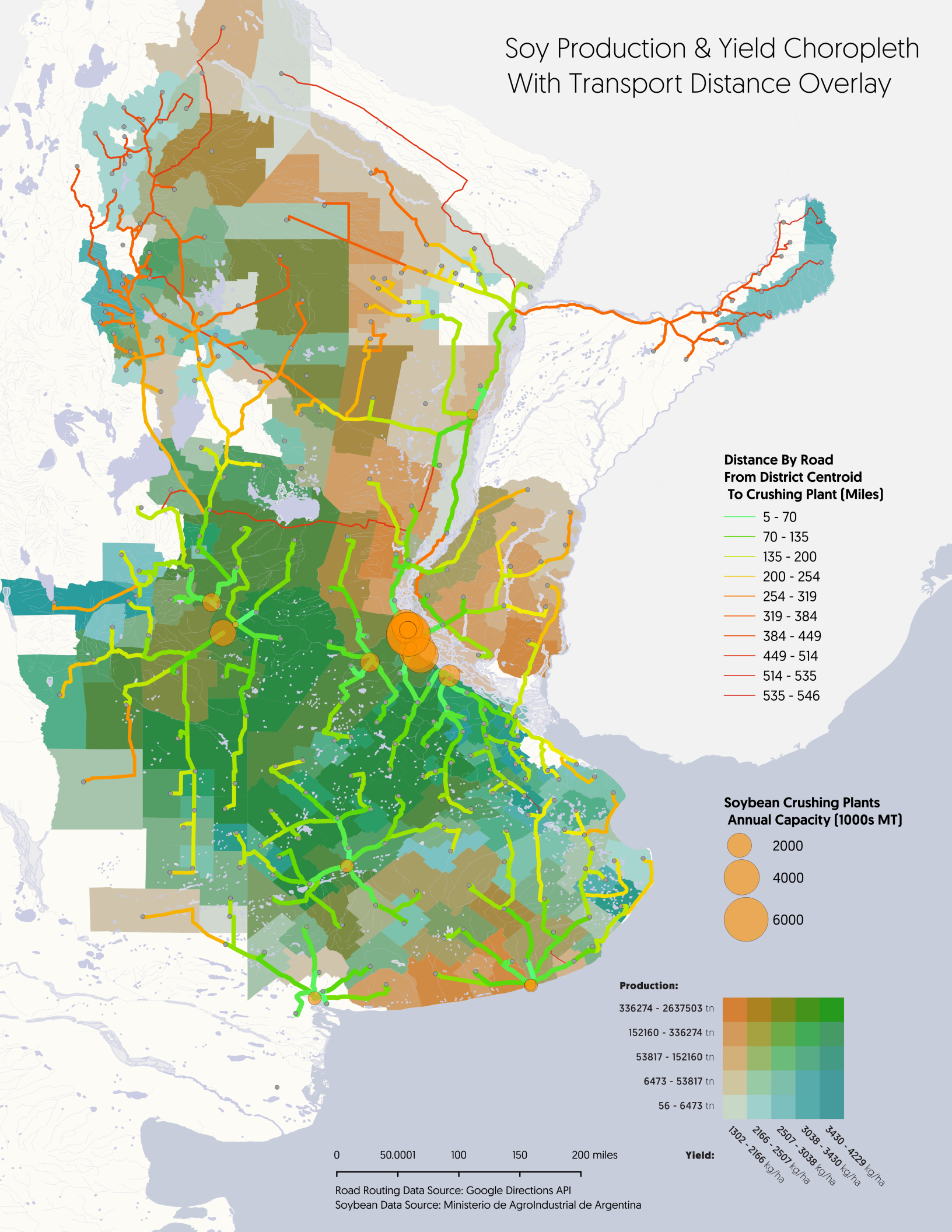


This produced the network shown on the left, and the map on the following page, which clearly shows that driving distance is highly correlated with production above and below what would be expected based on yield in all regions except the north, where it appears a soy crushing plant may be mising from the data.

Sources:

Visor IDE, Ministerio de AgroIndustrial de Argentina, ide.agroindustria.gob.ar/visor
Global Administrative Areas, Robert Hijmans, www.gadm.org
Google Directions API, developers.google.com/maps/
Instituto Geográfico Nacional, Centro Poblados, www.ign.gob.ar/NuestasActividades/sigign

Soy Production & Yield Choropleth With Transport Distance Overlay



Road Routing Data Source: Google Directions API
Soybean Data Source: Ministerio de AgroIndustrial de Argentina

Marginal Impact of Transportation Infrastructure on Agriculture

The data collected and presented in my previous three papers examine atributes of soy, from yield and production, to processing, and finally to the effect of transportation infrastructure. This paper seeks to establish an economic model for understanding those factors and ultimately generalize that model to predict the impact of changes in transportation infrastructure across the soy, corn, and wheat.

Spatial Model

To understand the significance to farmers of transportation to processing facilities, it is useful to model farmers' demand for processing capacity as a function of their yield rate and the distance to the crushing facility.

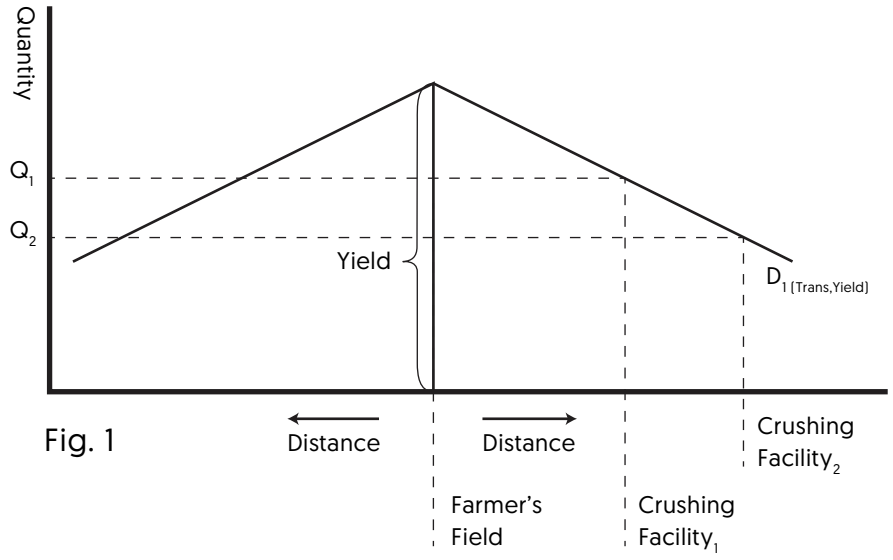


Fig. 1

Figure 1 on the previous page depicts a farmer's yield as the height of the central line—the amount of grain they can grow for a given amount of land—and the farmer's transportation costs as a linear function with a negative slope. The *y* axis represents the quantity of soybean processing the farmer demands, which in turn is a proxy for the amount of soy the farmer decides to plant.

This graph shows how farmers' decisions to plant soy would change based on the two variables considered. For example, an increase in a farmer's yield—from adopting a new technology, for example—would result in them producing more soy and therefore demanding more processing from plants at every distance:

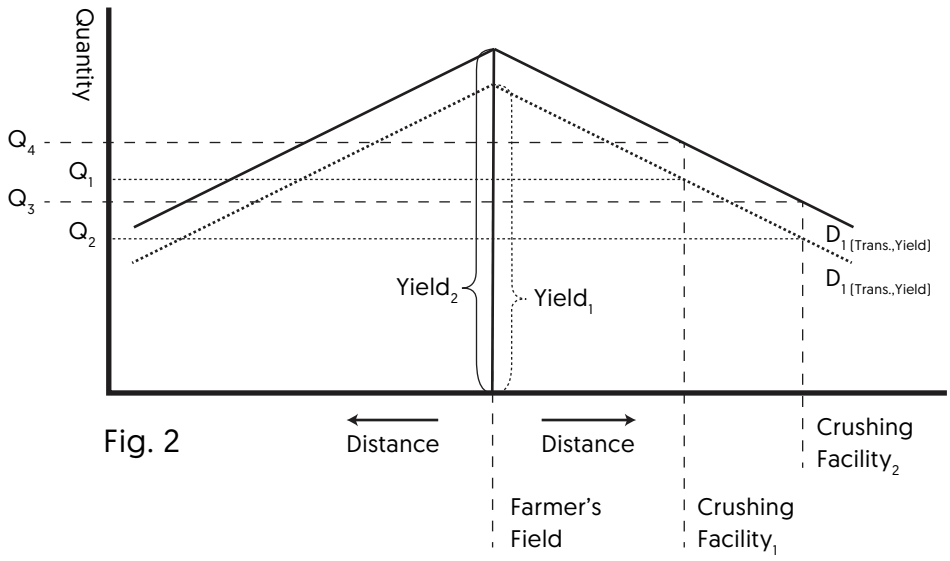


Fig. 2

This kind of shift would occur from the adoption of technologies such as Roundup Ready soy, enhanced fertilization methods, or other technological changes that happen locally at the farm. It is also possible to model a change in transportation costs, whether from improved roads, improved vehicles, less expensive fuel, or other factors. This kind of change would change the slope of the demand curve: as rate of transportation cost falls,

the demand curves will become less steep because farmers' decisions are less sensitive to the distance they need to travel to process their crop. Ceterus paribus, a decrease in transportation costs per mile will result in a greater demand for soybean processing at every distance.

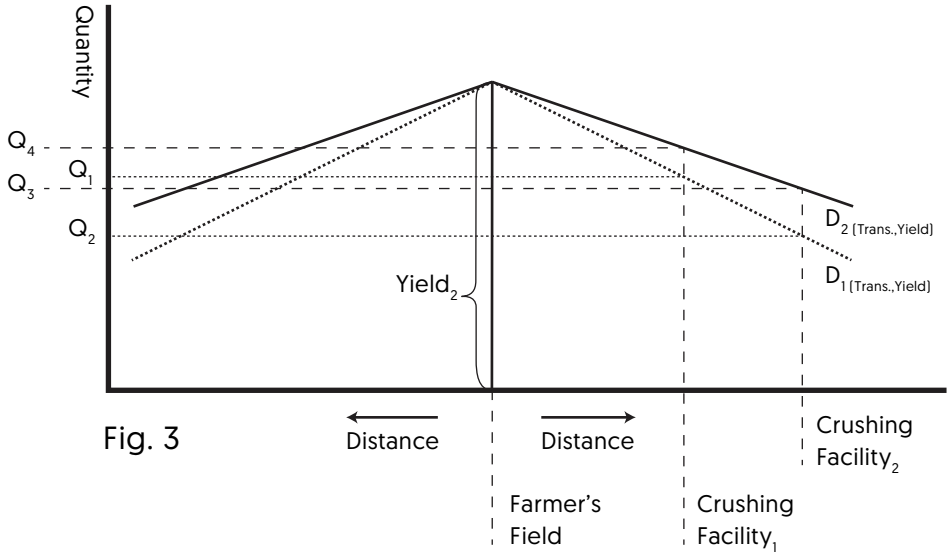
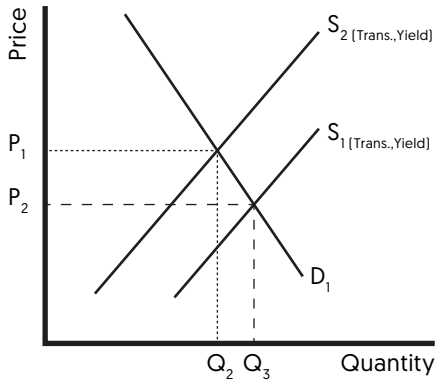


Fig. 3

This mechanism whereby the yield rates and transportation cost per mile faced by farmers changes the quantity of processing farmers demand then leads to shifts in price because quantity is changing. In the situations illustrated by figures 2 and 3, the result would be an increase in quantity to market and therefore a decrease in overall price, all else equal:



Opportunity Cost

Farmers have the opportunity to plant more than just soy, however. As a result, any planting decision will be made not as a function of the profitability of a single crop, but instead by analyzing the opportunity costs of planting one crop over another. Fortunately, the three primary grain crops planted in Argentina lend themselves to a similar transportation and yield model. While soy is transported to soybean crushing plants, corn and wheat are brought to markets such as the BCR Cash-Grain market to be sold, which can be modeled by a similar many-to-few network analysis along the roadways. Therefore, it should be possible to model farmers' decisions as a comparison between the expected profit for each crop where profit is the product of the futures price at planting time and the expected quantity harvested of each crop, and the quantity planted is equal to the quantity demanded of processing—whether that is crushing or listing on a market floor—and the quantity of processing demanded is in turn a function of each farmers historical yield and expected transportation costs for each crop.

Infrastructure Implications

This model is useful because it offers a mechanism to determine econometrically what impact a change in infrastructure could be expected to have on the quantity planted of each crop and therefore the price when those crops hit the market. This could provide infrastructure planners with a highly granular view of which improvements—potentially at the level of considering individual roads—could have the most impact on both farmers and end consumers, and allow planners to maximize the impact of their spending accordingly.

Summary

These four papers attempt to lay a foundation for an econometric analysis of crop planting decisions in terms of infrastructure, with the goal of creating a model that can advise and maximize real-world infrastructure building decisions and their effects on soy, wheat, and corn markets in Argentina.

Next Steps

The data analysis in this paper so far contains several obvious areas for improvement. Namely, some of the data are outdated: I need a better source data for soybean crushing plants, there are some lines in the network analysis that imply there might be flaws in the calculation methodology, and I would prefer to re-design my methodology to use OSM data instead of Google Maps data to make it easier to access more detailed information about the routes from each district to each crop purchaser. Furthermore, I need to develop a method for weighting the different plant capacities to make the selection of the optimal processing plant for any district more realistic.

Soy Planting Econometric Model

The maps and hypothetical models in the previous four essays lay a foundation for an econometric analysis of the way the road network in Argentina influences farmers' decisions to plant more or less soy, but does not lay out a specific model. The situation is significally complicated by the fact that every farmer is potentially considering 23 possible crushing facilities when deciding where to ship soy, each of which is a certain distance away and has a given capacity. One possible simplification of the problem–which did in fact give promising results as an early proof-of-concept—is to only consider the nearest crushing facility. However, this is not a safe assumption because there is a wide variety of types of facilities with wildly different capacities, and they are distributed incredibly unevenly throughout the country. A model studying planting decisions as a function of ease of access to processing facilities should not return similar values for one farmer who lives near a single small, low-capacity plant and a farmer who lives just outside Rosario, spoilt for choice near seven or eight of the largest capacity plants in the country; therefore every plant must be considered from the viewpoint of every district, as a function of both its capacity and distance from the district.

Though the ultimate goal of this research is to determine the impact of the road network on price of soy, and therefore determine the return on infrastructure investment from the perspective both of soy farmers and the overall agricultural industry, this first econometric model simplifies the problem by predicting the share of soy planted relative

to corn and wheat. The acerage planted was chosen for this analysis (as opposed to the area or quantity harvested) because it is a more direct indicator of the farmers' intent and is less likely to be effected by factors such as unexpected rainfall, drought, or other weather emergencies such as hail, flooding, or early frost. Therefore, the formula for the dependent variable is as follows, where S stands for siembra, the amount of each crop planted.

$$\text{Soyshare} = \frac{S_{soy}}{S_{soy} + S_{wheat} + S_{corn}}$$

This proportion also indirectly allows the model to adjust for the proportion of the department that is arable and actively used as farmland.

Independent Varianbles

The first three predictors of soyshare are quite obvious; the yields per hectare of soy, wheat, and corn. These are critical because it is one of the first things that comes to predicting what a farmer is likely to plant; it is a huge factor is the regional specialization of crops seen all over the world. In addition to these yields, another factor that could confuse and hide the effects of the transportation network are local laws and regulations, which vary from region to region and change the cost of everything from raw materials to the transportation of the finished product. Most of this regional variation occurs at the provincial level, so thirteen dummy variables have been included to isolate these provincial differences and separate them from the distance effects.

Finally, the most interesting effects should be captured by the log of 24 hour crushing capacity, the log of distance traveled, and the interaction between distance and capacity, which should provide insight in to how much farmers care about having more options even

if those options are further away. The regression was estimated in eViews, Estima RATS, and Python Statsmodels (for better integration with GIS and for mapping residuals) ordinary least squares implementations, with virtually identical results.

ESTIMA RATS Regression Output:

```
Linear Regression - Estimation by Least Squares
Dependent Variable SOYSHARE
Usable Observations                      6555
Degrees of Freedom                       6536
Skipped/Missing (from 8694)              2139
Centered R^2                        0.7074379
R-Bar^2                             0.7066322
Uncentered R^2                      0.9352554
Mean of Dependent Variable         58.098869205
Std Error of Dependent Variable    30.974829613
Standard Error of Estimate         16.777032294
Sum of Squared Residuals          1839680.1591
Regression F(18,6536)                 878.0310
Significance Level of F             0.0000000
Log Likelihood                    -27776.7993
Durbin-Watson Statistic                0.0970

    Variable                        Coeff      Std Error      T-Stat      Signif
********************************************************************************
1.  Constant                    -75.36872131  40.52215270     -1.85994  0.06293909
2.  SOJAR                         0.02130822   0.00026503     80.39822  0.00000000
3.  MAIZR                        -0.00154851   0.00012662    -12.22982  0.00000000
4.  TRIGOR                       -0.00105534   0.00022054     -4.78527  0.00000175
5.  LCP24H                       16.66035363   4.88739248      3.40884  0.00065633
6.  LDIST                         7.26112641   3.04935517      2.38120  0.01728478
7.  LCP_DIST                     -1.27966536   0.36818842     -3.47557  0.00051308
8.  PROVINCE(1)                  -1.98036472   1.16162851     -1.70482  0.08827597
9.  PROVINCE(2)                   6.31837973   2.24033122      2.82029  0.00481251
10. PROVINCE(3)                  16.33604631   1.13730343     14.36384  0.00000000
11. PROVINCE(4)                  19.34697680   1.28839284     15.01636  0.00000000
12. PROVINCE(5)                  14.51572082   1.48267844      9.79020  0.00000000
13. PROVINCE(6)                   0.00000000   0.00000000      0.00000  0.00000000
14. PROVINCE(7)                  -6.75265282   1.65170175     -4.08830  0.00004398
15. PROVINCE(8)                  -6.33585474   1.33115052     -4.75968  0.00000198
16. PROVINCE(9)                 -37.29618473   1.25009324    -29.83472  0.00000000
17. PROVINCE(10)                 -4.55791847   1.25708983     -3.62577  0.00029029
18. PROVINCE(11)                 10.66443211   1.58530397      6.72706  0.00000000
19. PROVINCE(12)                  6.39199116   1.40045084      4.56424  0.00000510
20. PROVINCE(13)                 -4.90431316   1.20021343     -4.08620  0.00004437
```

Python / Statsmodels OLS output (this is the model used to generate the residuals maps)

OLS Regression Results

Dep. Variable:	soyshare	R-squared:	0.707
Model:	OLS	Adj. R-squared:	0.707
Method:	Least Squares	F-statistic:	1089.
Date:	Tue, 21 May 2019	Prob (F-statistic):	0.00
Time:	05:44:38	Log-Likelihood:	-27777.
No. Observations:	6555	AIC:	5.559e+04
Df Residuals:	6536	BIC:	5.572e+04
Df Model:	18		
Covariance Type:	HC3		

	coef	std err	z	P>|z|
Intercept	-77.3491	38.996	-1.984	0.047
provincia[T.CATAMARCA]	8.2987	1.707	4.861	0.000
provincia[T.CHACO]	18.3164	1.134	16.152	0.000
provincia[T.CORDOBA]	21.3273	0.906	23.534	0.000
provincia[T.ENTRE RIOS]	16.4961	0.584	28.235	0.000
provincia[T.FORMOSA]	1.404e-13	6.94e-14	2.023	0.043
provincia[T.JUJUY]	-4.7723	1.979	-2.411	0.016
provincia[T.LA PAMPA]	-4.3555	1.064	-4.093	0.000
provincia[T.MISIONES]	-35.3158	1.639	-21.546	0.000
provincia[T.SALTA]	-2.5776	1.511	-1.706	0.088
provincia[T.SAN LUIS]	12.6448	1.413	8.950	0.000
provincia[T.SANTA FE]	8.3724	0.667	12.560	0.000
provincia[T.SANTIAGO DEL ESTERO]	-2.9239	1.526	-1.917	0.055
provincia[T.TUCUMAN]	1.9804	1.243	1.593	0.111
sojar	0.0213	0.000	75.481	0.000
maizr	-0.0015	0.000	-11.363	0.000
trigor	-0.0011	0.000	-4.880	0.000
np.log(CP24H)	16.6604	4.639	3.591	0.000
np.log(distance)	7.2611	2.995	2.424	0.015
np.log(CP24H):np.log(distance)	-1.2797	0.357	-3.584	0.000

Omnibus:	435.221	Durbin-Watson:	0.102
Prob(Omnibus):	0.000	Jarque-Bera (JB):	554.483
Skew:	-0.620	Prob(JB):	3.94e-121
Kurtosis:	3.702	Cond. No.	4.41e+20

These results, though they do not go the full distance to predict price, are incredibly promising. First of all, the signs on every variable make sense: More soy yield makes farmers more likely to plant soy, more corn or wheat yield makes them less likely, and all three of those variables are highly significant as expected. Also as expected, there are fairly large differences between provinces for other exognous reasons, and these are all significant as well with the exception of Santiago Del Estero and Tucuman, which are worth revisiting when examining the residuals in the next section.

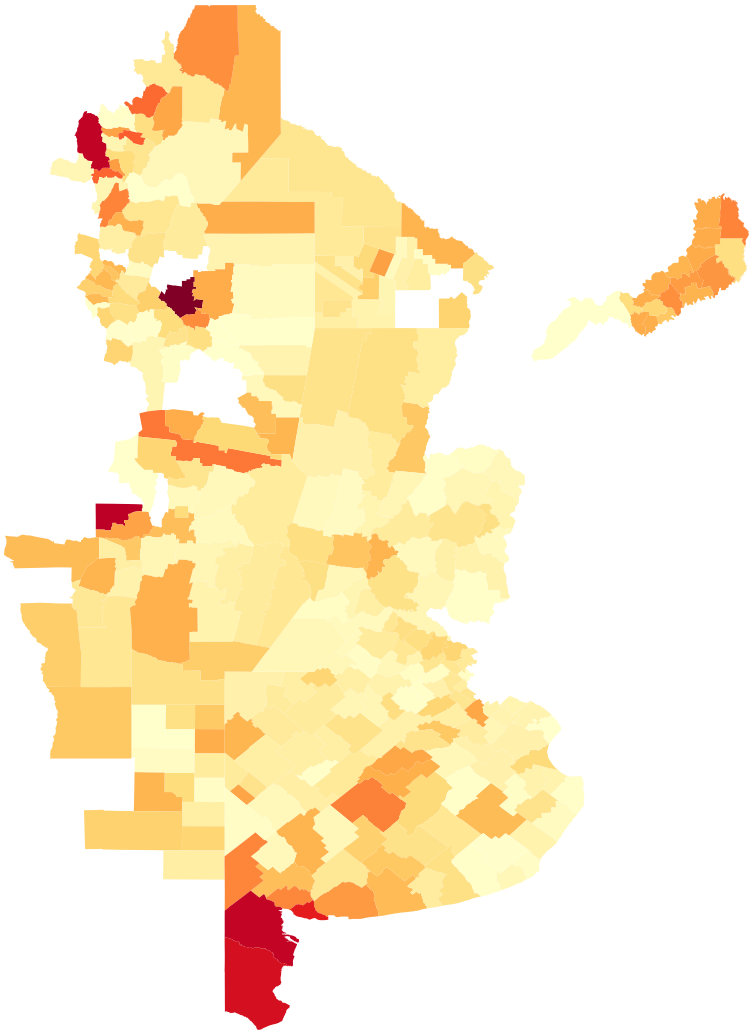
Interpreting the directionality and magnitude of capacity and driving distance is a little bit more complicated due to the interaction term. Considering just the portion of the equation regarding those three variables looks like this, where C is the log of 24 hour capacity and D is the log of distance:

$$\text{Soyshare} = \ldots \beta_{17}(C) + \beta_{18}(D) + \beta_{19}(C \times D)$$

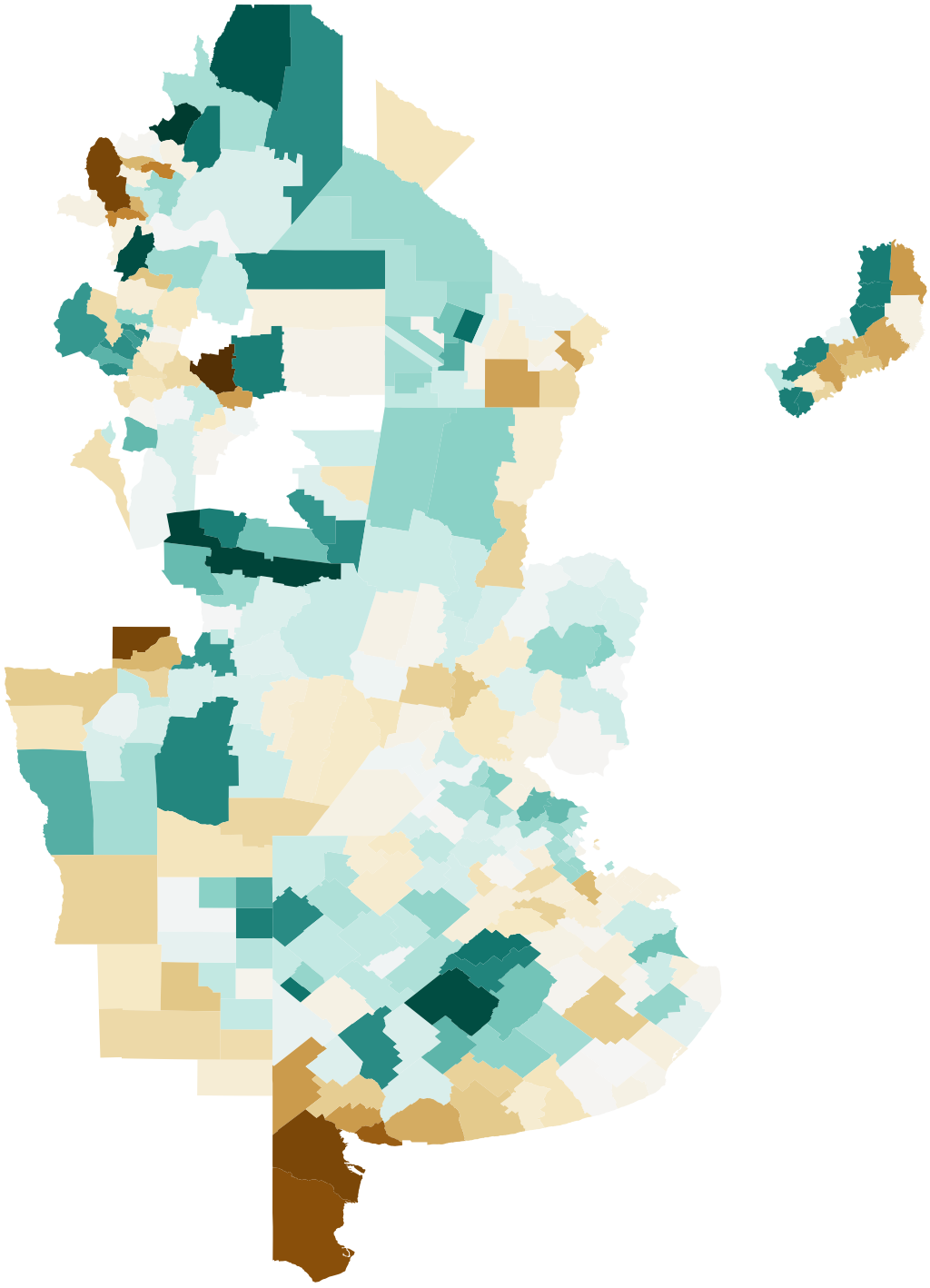
This explains the positive sign on β_{18}. Though it would initially appear that this model is finding that farmers prefer being further from crushing capacity, once the interaction is calculated the overall effect is negative. Solve for various values of C and D, and this model predicts farmers are less influenced by crushing capacity that is further away, just like hypothesized.

Residuals

Because this model is geospatial, interpretation of the residuals is more difficult than with a time series model where you are checking for correlation and missed effects along one dimension—forward and backward in time—to look for systematic undershooting and overshooting in the model or other patterns in the error terms. In a geospatial model, however, systematic patterns in the error term are distributed across two dimensions and not necessarily distributed on an even plane; there will always be effects from terrain and other physical characteristics that are too difficult to quantify in a model. Here is the magnitude of residuals from the Python/Statsmodels OLS estimation, plotted on a map of Argentina's agricultural districts, where red indicates a larger residual:

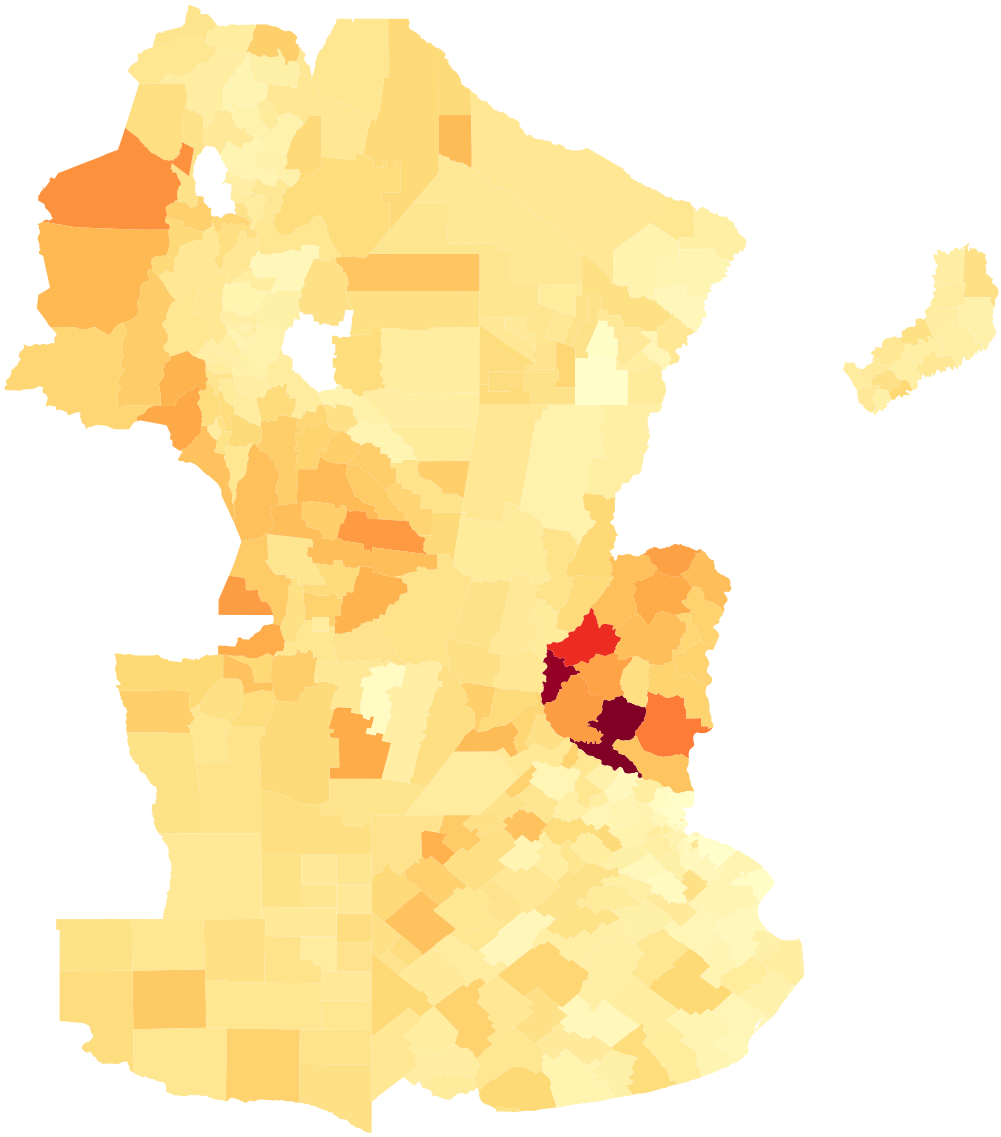


Potentially more interesting, however, is looking at the residuals with the sign included to look for patterns where the model might be systematically overshooting or undershooting, or look for unusual outliers. This plot shows areas were the model overestimates soyshare in brown, and underestimates in blue.

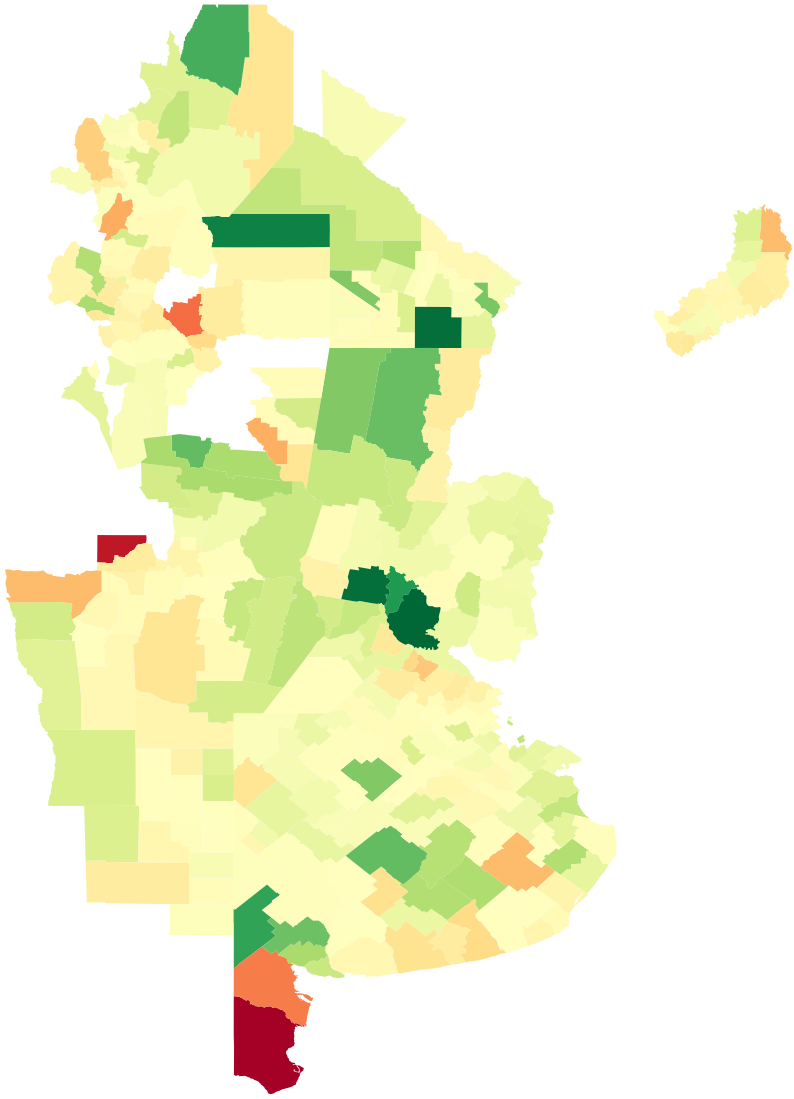


The model on the previous page makes it easy to identify there are six departments where the model is wildly overestimating, and all but one of them are on or near the border with Chile. In all but one of these districts the actual soyshare is near zero, and in all cases the model is predicting about 50 percentage points higher than the actual value. The soyshare values are so anomously low that this could be due to data collection or other issues, or perhaps interactions across the border. Re-estimating the model without these outlier districts produces a 0.05 increase in R squared values, but reduces significance in several exogenous variables, so it is worth investigating why they are so different. More interesting, perhaps, are the areas in blue in the south-west part of Buenos Aires province—it seems there might be an additional factor making this region so attractive for soy farmers, even though it has roughly the same crusher capacity access as córdoba province to the north.

Another way to consider the fit of this model is to look at the inefficiency of the road network, calculated as the mean of the multiples of the road distance to each crusher relative to the straight line distance. Though it sounds straightforward, it is important to take in to consideration the map projection when performing a calculation like this, so the data were first reprojected to EPSG:22171, an equidistant projection centered at the Posiciones Geodesicas Argentinas 1994 datum which allows easy calculation of distance in meters, the same units and used in the google driving data. After transforming both start and end points to EPSG:22171, the distance is calculated along the WGS84 ellipsoid so that it will match the driving directions calculations. This road network inefficiency is shown on the next page.



This map clearly highlights the areas discussed in essay, especially the districts directly across rio paraná from rosario; traffic from those districts must flow through just three bridges, making it a significant detour from straight-line distance. Because these districts have less direct road systems than their neighbors, this is where the roads model should have the greatest effect; that can be seen on the next page with a plot showing the difference of squared residuals between this model and a restricted model using only the yield rates–the correlation between these maps is clear.



In conclusion, this model establishes that the data are sufficient to demonstrate directionality of all effects, and proves that the road network plays a significant role in Argentina's soy market. The analysis in these papers so far contains several obvious areas for improvement, primarily in finding a better source data for soybean crushing plants, and making sure the planting data is as aligned as possible with the road network, given the fact that the road network has no single date where it is current because it is always being updated.