Commodity Trade Matters
Data Description

Thibault Fally and James Sayre
UC Berkeley ARE

August 2018

Below we describe the sources and procedure used to generate the data in our paper. We believe that this dataset may be of use to other researchers studying commodity trade, so we provide our data online at http://are.berkeley.edu/ fally/data.html, and intend to keep this information updated. When assembling commodity statistics on production, prices, and trade, the data are often reported at different levels of aggregation, and so we describe the associated difficulties of this below. We attempt to aggregate these data to the most precise level possible, and provide correspondence tables between the various sources of data used in the paper.

Production data. The British Geological Survey (2015) provides world mineral production statistics at the country level from 1913 to 2015, which is the main source of mineral production data.[1] The production data can be found online at the BGS website and is provided by the Natural Environment Research Council. For many commodities, the information is organized at the commodity level, but provided at the "subcommodity" level. For instance, "Titanium" is reported as Struverite, Titanium slag, Ilmenite, Rutile, Leucoxene, and simply as Titanium. In many of these cases, we sum production at the subcommodity level up to the commodity level, however in some cases, we use this information to aggregate the production data to a different commodity.

The main source of agricultural production data is FAOSTAT, provided by the Statistics Division of the Food and Agriculture Organization of the United Nations (2017), which provides data from 1960 to 2014 on the production of primary and processed agricultural products at the country level, which is also used by Costinot and Donaldson (2012) and Costinot et al.

[1]From 1960 to 2015, this information is available in spreadsheet format, earlier years are available only in PDF format. The US Geological Survey also provides mineral production data at the country level, however we do not use this because, as to our knowledge, the data provided by USGS are available only for 2001-2014 in spreadsheet format. Where data are available to compare, in many cases, the USGS and BGS production data match, and when they don't, the differences are often minor. As it is difficult to say whether one source is more precise than the other, we prioritize the BGS production data.

(2016).[2] The FAO provides correspondence tables for conversion of its own product classification to the 1996 version of the Harmonized Classification system, which we then use to create a correspondence of our own to the HS 1992 nomenclature.

Supplementally, we employ production data from the Global Trade Analysis Project, or GTAP version 8 (Aguiar et al., 2012), which provides production (in terms of value) data at the industrial sector level by country for 2007. While these data is mostly used in the calibration to provide the output of downstream industries (such as Motor Vehicles, Electronic Equipment, etc.), in a few cases we use the data to provide information regarding the output of primary commodities. We use GTAP production statistics for unrefined sugar, paddy rice, wheat, coal, crude oil, and natural gas in our calibration for 2007.[3]

Trade data. Trade information comes from the BACI database, constructed by CEPII and based on UN-Comtrade data (Gaulier and Zignago, 2010), and provided at the 6-digit level of the Harmonized Commodity Description and Coding System (HS). We use the HS 1992 nomenclature, as it provides the longest series, covering the years 1995 to 2014 (as of writing). Since the commodity lithium is not classified in the HS 1992 nomenclature, we use HS 1996 data to provide trade information for lithium. In order to match production and trade data, we further aggregate the trade data to match the level of granularity in the production data.

Data Aggregation We provide online a correspondence table between our aggregation codes and trade data, in addition to providing production, price, and input-output data used in the paper. For all these scattered sources, we try to remain as close as possible to the Harmonized Classification System (HS). When aggregating directly to a six digit HS code is not possible, we use a simple notation. We use the letter "A" (potentially followed by several zeroes) to denote that all listed HS6 products starting with the numbers before "A" are aggregated into this code. For instance, the aggregation code 3104A0 (Potash) includes the six digit codes 310410, 310420, and 310430, and any other codes starting with 3104 (only 310490, in this case). The letter "X" indicates that the aggregate code contains a selection of HS six digit products. For instance, our aggregation code 0810X0 (Berries) includes the six digit HS codes 081020 and 081040, but not the six digit code 081010 (Strawberries). However, any code containing either "A" or "X" may also contain additional six digit HS codes, when the level of production data requires aggregation above the HS four digit level, which should be noted. In the cases where aggregation is required, we compute production value at the most disaggregated level (that is, the level that prices are provided at), and aggregate this value, rather than aggregating quantities. It is for this reason that the data we provide online are slightly more disaggregated than the data we use in our baseline calibration; we provide data at the level at which we can provide informative quantity information. We provide a correspondence between the more disaggregated data we provide online and our baseline specification online.

[2]FAOSTAT also provides information regarding the production of livestock and animal products which we do not use, as it is difficult to argue that livestock requires natural resources as concentrated as those required in the production of minerals and other agricultural products.

[3]We do not include GTAP production statistics in the data we provide online. For these commodities we provide the data supplied by the BGS and FAO, which seems to be similar, although somewhat less reliable for a few outliers, mainly developing countries.

Price data. The United States Geological Survey provides the Historical Statistics for Mineral and Material Commodities database (Kelly and Matos, 2014), which catalogs prices of mineral commodities in the United States from 1900 to the present, and is the most comprehensive source of yearly price data available for minerals. One shortcoming of the database is that it does not cover mineral prices for countries other than the US.[4] One potential option to address this is by using export unit values from trade data instead as a proxy for producer prices. This route has well known shortcomings: unit values are frequently noisy, we find very large ranges in these values across countries, and observe occasional massive yearly spikes in unit values not reflected in the USGS price data that seem unlikely. These issues are most pronounced for developing countries. Further, since the trade data must often be aggregated to match the production data, it is unclear whether the use of quantity information in such settings makes sense. Using unit values from the trade data is often problematic – resulting in many observations where the value of production of one or more commodities we observe exceeds GDP for the same time period. Reassuringly, we find that except for the aforementioned deviations and outliers, the USGS price data generally track fluctuations in unit values quite well, especially for large, developed countries.

One remaining difficulty is that the prices in the Mineral and Material Commodities database are for refined minerals, rather than for primary commodities such as ores. Therefore, using prices directly from the database would result in production values of minerals far higher than the actual value of production in those cases, especially for countries where refining of primary commodities produced domestically is done abroad. To address this, we "downscale" commodities based on United States export unit values, which generally look similar to the trends in the USGS price data.[5] A scaling factor, β, is chosen to minimize the sum of squared distance between the USGS price and the unit value price for a given commodity, so long as that scaling factor is less than one. To give a concrete example, to give a price to the production of Chromium Ore (the unrefined primary ore), we scale the price given for Smelted Chromium (a refined secondary product) by the US export unit value for Chromium Ores (HS code 261000), which results in assigning a price for producers of chromium ores as $\beta = .368$ times the price for refined Chromium. Since one would expect that changes in demand for processed metals affect demand for their primary ores in similar ways, this should imply that prices for primary commodities have similar trends, but lower overall levels. Indeed, looking at the US unit values for primary and processed mineral commodities for the small number of commodities we use this procedure on, this seems to be the case (in total, we perform this procedure for primary ores and unprocessed products of Asbestos, Aluminum, Antimony, Boron, Chromium, Cobalt, Copper, Gold, Iodine, Lead, Magnesite, Manganese, Molybdenum, Nickel, Silver, Tin, Tita-

[4]By applying world prices to mineral production throughout the world, we are essentially assuming that minerals are fully homogenous, or that the trade elasticity is very large. While this is certainly not accurate, it is a more plausible assumption for minerals than other traded goods (although many authors have found that the trade elasticity is generally not higher for agriculture or commodities as a whole, Caliendo and Parro (2015) find evidence of a higher trade elasticity for minerals and petroleum). Further, in the text we demonstrate that our results are less sensitive to magnifications of the trade elasticity than in standard models, and in our context, it seems unlikely that having country specific prices would alter the estimates for the gains from trade very much. In other contexts, this would likely be a larger limitation.

[5]We could downscale commodities using country specific scaling factors as well, but the concern again is how reliable unit values are for reporters that are developing countries.

nium, Tungsten, and Zinc). Of these commodities, there are only six commodities for which we need to aggregate trade data to match the level of production, avoiding concerns about the suitability of aggregating quantities of trade. For the remaining six (Beryl, Boron, Copper, Molybdenum, Platinum, Rare Earth Minerals), we find that unit values from exports still follow the USGS prices closely. Figures 1 plot the comparison of US export prices and USGS prices per ton for a selection of commodities we perform this procedure on.

The USGS price data do not contain any information on uranium and fuels prices, so these data are complemented by the International Monetary Fund (IMF) Primary Commodity Price Series database for monthly uranium prices (which we aggregate up to yearly prices) (Commodities Team of the Research Department, IMF, 2017), the World Bank Commodity "Pink Sheets" for petroleum and coal prices (World Bank Group, 2017), and data from the U.S. Energy Information Association (2017) (EIA) on the producer (wellhead) price of natural gas, all of which are in current US dollars.

For agricultural products, FAOSTAT provides yearly country-level agricultural price data. This information is listed at the same level as the production data, and only aggregate these data after computing the production value of each commodity at level of aggregation the FAO provides. Although the FAO provides price information for many commodities in terms of current US dollars, often the prices are provided in terms of local currency units. When available, we prioritize the prices as listed in terms of US dollars, supplemented by an exchange rate table for each country provided by the IMF-IFS database. Many commodities listed in the FAOSTAT are missing country level price information, for which we replace with the world median price.[6] In some cases, the producer price of a given commodity in one country can be almost 1,000 times as large as the median world price. These cases seem highly unlikely to reflect prices that producers would receive on the world market, and strongly inflate the value of production of these commodities, resulting in cases where the production value of a commodity exceeds reported GDP. Therefore, we omit country price data for commodities that are 50 times greater than the median world price, replacing these cases with the median world price.[7]

Commodity end use. GTAP provides information on the use of broad commodity sectors by downstream industrial sectors. We employ GTAP information to provide country level end-usage data for agricultural commodities and fuel products. However, as GTAP aggregates mineral commodities into only 2 categories, we combine it with USGS end-use data (Barry et al., 2015) for minerals. The USGS end-use data provide information on the relative use of mineral commodities by NAICS industry in the United States. We then match each NAICS code to the GTAP industrial classification system manually, and use this to match each commodity to the intensity of usage by each downstream GTAP industrial sector. Occasionally, the USGS data do not provide the relative frequency of mineral end-use by NAICS downstream sector for some commodities. However, the USGS still provides information on the NAICS downstream sectors that use the commodity, just not the relative proportions across industries. In these cases, we

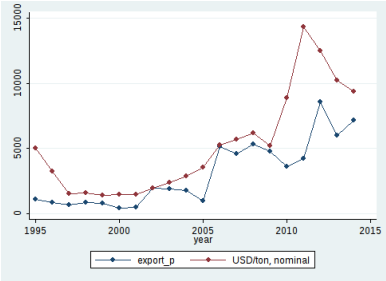
[6]We use the median price because in several cases there are outlying prices that bias the prices strongly upward.

[7]We have also tried replacing world prices with regional averages, however unfortunately in some regions there may be only one price, so averaging will bias all prices for a region upwards.

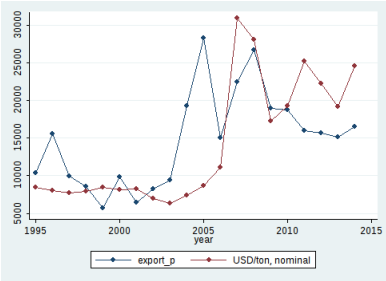
use the relative end use frequencies across downstream sectors for the respective commodity category from GTAP, but renormalize these frequencies by removing downstream industries not mentioned as using the commodity by the USGS. In the case of three commodities in our baseline calibration, there is more than one end use table for each "commodity" we use. For instance, "Platinum Group Metals" uses end use tables for Platinum, Palladium, Rhodium, and Iridium; "Vermiculite" uses end use tables for Vermiculite and Perlite, and "Niobium et al." uses end use tables for Vanadium and Tantalum. In such cases, we take a weighted average of these respective end use tables, where the weights are computed as the worldwide production value in 2007 for each end use mineral over the value of all constituent minerals in a commodity. This results in zero weights for Vermiculite, Rhodium, and Iridium, within "Niobium et al." the Vanadium end use table receives a weight of 0.84 and the Tantalum table has a weight of 0.16. Within "Platinum Group Metals", Platinum receives a weight of 0.6, Palladium receives a weight of 0.4, the remaining minerals have zero weights since they have zero production value in 2007.

Other Data Additionally, for our simulations, we employ GDP, natural resource rents, and value added data provided by The World Bank (2017).

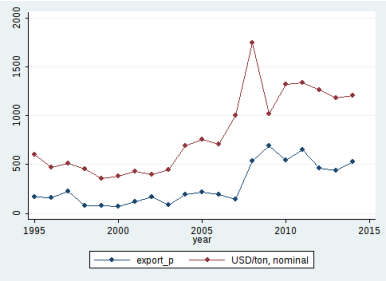
Figure 1: Comparison of USGS prices and US export prices (Red line is USGS provided price, blue is US export unit value, in USD per ton)



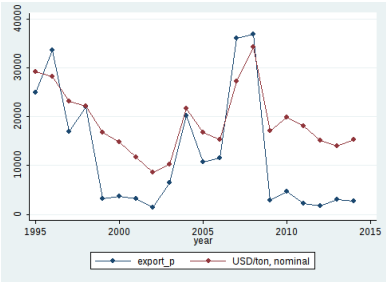
(a) Antimony



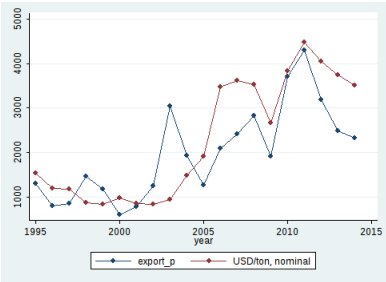
(b) Bismuth



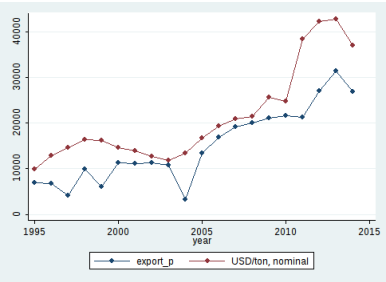
(c) Chromium



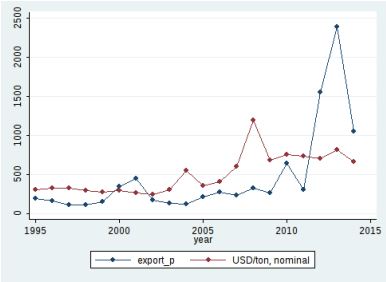
(d) Cobalt



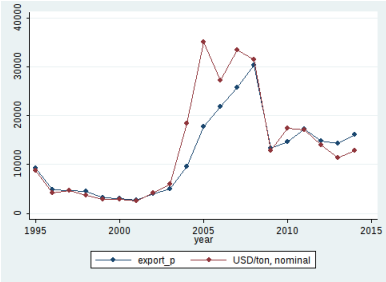
(e) Copper



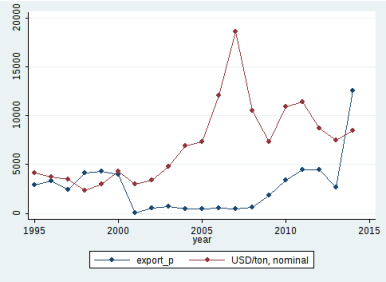
(f) Iodine



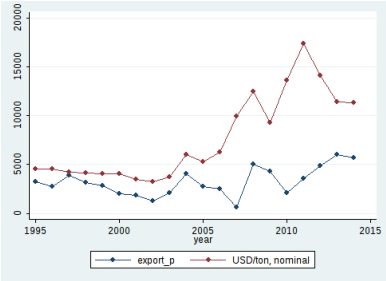
(g) Manganese



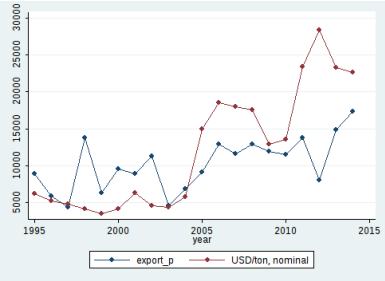
(h) Molybdenum



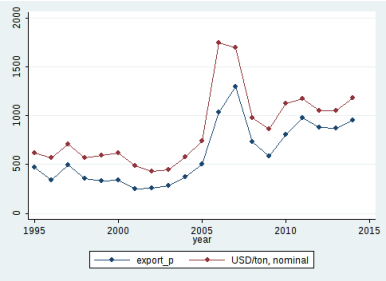
(i) Nickel



(j) Tin



(k) Tungsten



(l) Zinc

Using gravity to fill in zeros in autarky counterfactuals

In section **??**, we describe the issues presented when a country has positive demand for a commodity but no domestic production for measuring the gains from trade when considering full movements back to autarky. To partially address these concerns, we use predicted bilateral trade and production instead for autarky counterfactuals. Ideally, we would estimate the following equation for each commodity using PPML:

$$\begin{aligned}\log X_{nig} = \; & FX_{ig} + FM_{ng} + \beta_{Dist,g} \log Dist_{ni} + \beta_{Contig,g} Contig_{ni} \\ & + \beta_{Lang,g} CommonLang_{ni} + \beta_{Colony,g} Colony_{ni} + \beta_{HomeBias,n,g} \mathbb{I}(n=i) + \varepsilon_{nig},\end{aligned} \tag{1}$$

and then use the predicted trade flows $\widehat{X_{nig}}^{pred}$ to provide us with predicted production for each commodity, defined as $\widehat{Y_{ig}}^{pred} \equiv \sum_n \widehat{X_{nig}}^{pred}$. However the home bias, that is, the estimated log increase in trade flows due to moving inside a country's borders, is not identified if internal flows are treated as missing.

A first solution would be to impose the home bias effect to be uniform across countries and estimate it using countries for which internal trade data are not missing. However, this would lead to overstatement of the home bias effect because of a selection bias. Countries with reported production data are more likely to be among the largest producers, and thus mechanically are more likely to consume more of their own domestic output. This induces an upward bias in the border effect coefficient, and results in predicted internal trade flows that are often implausibly large.

The solution that we propose involves two steps. First we estimate equation (1) with available trade flows. An important property to note is that the sum of fitted external flows for a country equals the sum of its observed exports or imports for that country, a property specific to PPML, with the inclusion of exporter and importer fixed effects (Fally, 2015). The same holds for fitted internal flows, which equal observed internal flows in each country where internal flows are not missing, as long as country-commodity specific border effects are included in the regression. Therefore, with missing internal flows, we can use equation (1) to predict these flows up to the home bias coefficient $\beta_{HomeBias,n,g}$ for that country. We denote such fitted flows by $\widehat{X_{nng}}(\beta_{HomeBias,n,g})$.

In a second step, to estimate the home bias coefficient when internal flows are missing, we employ GTAP data at a more disaggregated level (which features almost no missing internal flows), and assume that the home bias coefficient is uniform within the country and GTAP sector G in which the commodity $g \in G$ belongs: $\beta_{HomeBias,n,g} = \beta_{HomeBias,n,G}$. We then calibrate the home bias such that predicted internal flows are equal to observed internal flows for the GTAP sector in that country. Using adding-up properties of PPML, this is equivalent to calibrating the home bias coefficient as:

$$\hat{\beta}_{HomeBias,n,G} = \log\left(\frac{\sum_{g\in G}\widehat{X_{nng}}(0)}{X_{nnG}}\right) + \log\left(\frac{\sum_{k\neq n} X_{knG}}{\sum_{g\in G}\sum_{k\neq n} X_{kng}}\right)$$

where the numerator of the first term uses fitted flows constructed without the home bias coefficient ($\beta_{HomeBias,n,g} = 0$), and the denominator is observed internal trade for the aggregate GTAP sector. As a GTAP sector may also contain other goods not covered in our analysis, we

adjust our estimation for the share of such goods in the aggregate GTAP sector trade using the second term.

References

Aguiar, A., R. McDougall, and B. Narayanan (2012). *Global Trade, Assistance, and Production: The GTAP 8 Data Base*. Center for Global Trade Analysis, Purdue University.

Barry, J. J., G. R. Matos, and W. D. Menzie (2015). *A Crosswalk of Mineral Commodity End Uses and North American Industry Classification System (NAICS) codes*. US Geological Survey.

British Geological Survey (2015). World Mineral Statistics Archive.

Caliendo, L. and F. Parro (2015). Estimates of the Trade and Welfare Effects of NAFTA. *The Review of Economic Studies 82*(1), 1–44.

Commodities Team of the Research Department, IMF (2017). IMF Primary Commodity Prices.

Costinot, A. and D. Donaldson (2012). Ricardo's theory of comparative advantage: Old idea, new evidence. *American Economic Review 102*(3), 453–58.

Costinot, A., D. Donaldson, and C. Smith (2016). Evolving comparative advantage and the impact of climate change in agricultural markets: Evidence from 1.7 million fields around the world. *Journal of Political Economy 124*(1).

Fally, T. (2015). Structural gravity and fixed effects. *Journal of International Economics 97*(1), 76–85.

Food and Agriculture Organization of the United Nations (2017). FAOSTAT statistics database.

Gaulier, G. and S. Zignago (2010, 10). BACI: International Trade Database at the Product-Level. The 1994-2007 Version. *CEPII WP No. 2010-23*.

Kelly, T. and G. Matos (2014). Historical statistics for mineral and material commodities in the United States (2016 version). *U.S. Geological Survey Data Series 140*.

The World Bank (2017). World Development Indicators (1960-2016).

U.S. Energy Information Association (2017). Natural gas prices.

World Bank Group (2017). World Bank Commodity Price Data (Pink Sheets), Global Economic Monitor Commodities.