Commodity Trade Matters
Data and Code Description

Thibault Fally and James Sayre
UC Berkeley ARE

November 2017

1 Data and Code Description

This document describes the sources and program code used to generate the data used in our paper. We believe that this dataset may be of use to other researchers studying commodity trade, and so we have tried to make the data construction process as transparent as possible. One difficulty of assembling commodity statistics on production, prices, and trade, is that the data are often reported at different levels of aggregation. We attempt to aggregate this data to the most precise level possible, and provide a correspondence table between the various sources of data used in the paper. The data can be found online, and we intend to keep this information updated on a semi-frequent basis.

2 Data Description

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 production data. 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 for three reasons: first, the data is provided in a more difficult format to clean; in many cases the formatting of the production data varies year by year and by each commodity. Second, to our knowledge, the data provided by USGS is also available only for 2001-2014 in spreadsheet format. Finally, in many cases, the USGS and BGS production data match, and when they don't, the differences are often minor, it is difficult to say whether one source is more precise than the other. For these reasons, we prioritize the BGS production data, which 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 agricultural products at the country level. We use data from FAOSTAT on the production of primary crops and processed crops, but do not use any information on livestock or animal products, since it is difficult to argue that livestock requires natural resources as concentrated as those required in the production of minerals and other agricultural products. The FAO provides correspondence tables for conversion of its own product classification to the 1996 version of the Harmonized Classification system, which we then convert to the 1992 nomenclature.

Supplementally, we employ production data from the Global Trade Analysis Project, or GTAP (Aguiar et al., 2012), which provides production data at the downstream sector level by country. Currently, we are using GTAP version 8, which provides data for 2007.

Trade data. Trade information comes from the BACI database, constructed by CEPII and based on UN-Comtrade data (Gaulier and Zignago, 2010). The data are detailed 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). Worth noting is that 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. Additionally, UN-Comtrade provides historical trade data at the SITC classification going back to 1970. We perform a similar aggregation procedure for the trade information in the SITC format, however the level of aggregation is less refined, because the SITC classification defines many commodities broadly. Therefore, we expect our historical results to be more similar to other studies, which have defined commodities more broadly, such as "minerals".

Price data. The United States Geological Survey provides the Historical Statistics for Mineral and Material Commodities database (Kelly and Matos, 2014), which catalogs current prices of mineral commodities in the United States. The data range from 1900 to the present, and are the most comprehensive source of yearly price data available. However, one shortcoming is that the data not cover mineral prices for countries other than the US. 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, and we find large ranges in these values across countries, and unlikely spikes in unit values not reflected in the USGS price data, particularly 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 the 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, the USGS price data generally tracks fluctuations in unit values quite well.

One remaining difficulty is that the prices listed by the USGS are for processed commodities, rather than for primary commodities. 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. 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, Aluminium, Antimony, Boron, Chromium, Cobalt, Copper, Gold, Iodine, Lead, Magnesite, Maganese, Molybdenum, Nickel, Silver, Tin, Titanium, 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 the export prices follow the USGS prices closely. Figures 1 and 2 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 does not contain any information on uranium and fuels prices, so this data is 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 (GEM Commodities, 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 this 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, but supplementally use a exchange rate table for each country provided by the IMF International Financial Statistics database. Unfortunately, many commodities listed in the FAOSTAT are missing country level price information. To generate prices for these cases, we use the world median price. The reason for use of the median price is that in several cases, there are outlying prices that bias the prices strongly upward. 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. Therefore, we omit country price data for commodities that are 50 times greater in price than the median world price, replacing these omitted prices with the world median price. We have also tried replacing world prices with regional averages, however unfortunately in some regions for some commodities, there may be only one price, and if that price is much higher than the world median, it will bias all prices for a region upwards.

Other Data For calibrating our gravity equations, we use bilateral distance and geography data from CEPII (Mayer and Zignago, 2011). Additionally, for our simulations, we employ GDP, natural resource rents, and value added data provided by The World Bank (2017).

Commodity end use. GTAP provides information on input-output linkages between categories of primary commodities and broad secondary and tertiary industrial sectors. We employ GTAP information to provide end-usage data for agricultural commodities and fuel products. Since GTAP aggregates mineral commodities into only 2 categories, we employ USGS end usage data (Barry et al., 2015). The USGS end usage data provides information on the percentage each downstream NAICS industry uses a given mineral. 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 does 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 use the relative end use frequencies across downstream sectors for minerals from GTAP, but renormalize these frequencies by removing downstream industries not mentioned as using the commodity by the USGS.

3 Code Description

The data for this paper is constructed using both Stata and Python. The Python scripts are written to either compile correspondence tables, produce descriptive statistics that require looping over .tex files, or download data from various sources. The Stata scripts are written to clean and compile the data, and run most of the analysis. Found online, Stata do-file masterscript.do outlines the order in which our programs must be run to compile the data used in this paper, including any scripts written in Python. Additionally, all of the Stata do-files can be run within this file, which allows the user to have to only adjust the directories to the ones on their local computer once. Although this file provides a precise list of programs to run in order, here we simply provide a rough outline of the functions and order of the programs used in our analysis and data construction.

Data Aggregation Found online, we provide a correspondence table between the various sources of price, production, trade, input-output data and aggregation codes. We attempt to provide as close as a correspondence between these scattered sources of data and the Harmo-

nized Classification System[1]. When aggregating directly to a six digit 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 this document, the first column ("code") provides the HS6-like code used in most cases to aggregate each source of data. However, in several cases, it is necessary to aggregate the data further, but still use the individual prices, which are provided at a lower level of aggregation. In these cases, a 1 in the column "aggregate" indicates that the data should be aggregated into an HS6-like code given by the column "aggcode". In these cases, the program code aggregates the production and price data for minerals into the code given by the "code" column, computes the value of production (as opposed to the quantities of production), and then aggregates the value of production into the code given by the "aggcode" column. A similar procedure occurs with the agricultural production data, except the production value is computed at the level at which the FAO provides price data, and then aggregated into either the aggregation code provided by "code" or "aggcode" columns (depending on the "aggregate" column indicator). The "included" column indicates whether a commodity is currently included in our analysis (there are several commodities which do not have price information for the duration of the period for which BACI trade data are available (1995-2014), which we exclude). The remaining columns provide information on how to match each aggregation code to price, production, and trade data. Of note are the "min_prod_units" and "min_price_ratio" columns, which provide the conversion factor required to convert units of production into tons in some cases, and *either* the aforementioned US export price deflation factor for several primary mineral commodities, or in the case of Coal, Natural Gas, and Petroleum, a conversion factor to convert prices into US dollars per ton. Also found within this table are GTAP industrial codes for each commodity, the commodity type, and the BEC index of production sophisication, which is used to compute the level of processing, which takes 1 for primary products, and 2 for processed products. The python script "buildmastercorrespondence" then takes this correspondence file, and uses it to generate smaller correspondence tables, used by each script responsible for cleaning the production, price, and trade data.

Data Assembly The stata do-file "BuildProduction" uses the aggregation table provided by the previous script to aggregate BGS mineral and fuel production data to the desired level, adding country codes at the same time. The python script "FAOCleaning" combines agricultural production and price data from FAOSTAT, converts FAO codes into Harmonized System codes using a provided correspondence table, converts local currency units into current US dollars (when necessary), and computes world average prices for each commodity, when a

[1]Additionally, with historical data, a correspondence between our varied sources of data and the SITC classification system.

country level price is not available. The stata do-file "build_trade_data.do" aggregates BACI trade data using the correspondence table computed earlier, and later computes export/import unit values for these products at the country level. The stata do-file "CleanPrices" cleans and assembles mineral and fuel prices using another correspondence table, provided by the script "buildmastercorrespondence". The stata do-file "build_cntry_prod_data" gathers the data compiled by the aforementioned scripts, computes production values for each commodity, then aggregates this data to the upper (final) level of aggregation. This script also computes the total value of country-level exports and imports for each product.

Data Analysis Various other scripts use this data to compute other parts of our analysis. The stata do-file "gravityequations" runs the gravity equations estimations and produces simulated production and trade flow data. The stata do-file and python scripts named "StylizedStatistics" compute many of the stylized facts found within the paper. The stata do-files "EstimateVolatility" and "EstimateElasticities" estimate price volatility and use counterfactual "China Shock" data to estimate price elasticities for the commodities in our sample. The R script "makeproductionmaps" generates world maps for production statistics of agricultural products and the distribution of world gains from trade.

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. Technical report, US Geological Survey.

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

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

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 Working Paper No. 2010-23*.

GEM Commodities, World Bank Group (2017). World Bank Commodity Price Data (Pink Sheets).

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*.

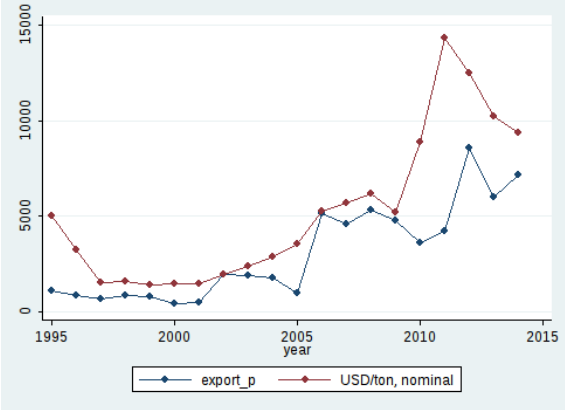
Mayer, T. and S. Zignago (2011). Notes on CEPII's distances measures: the GeoDist Database. *CEPII Working Paper 2011-25*.

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

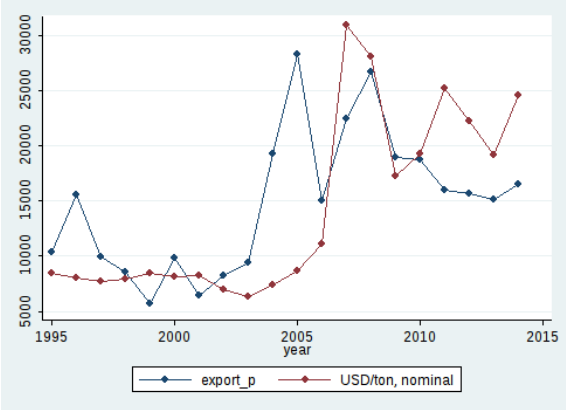
U.S. Energy Information Association (2017). Natural gas prices.

4 Tables

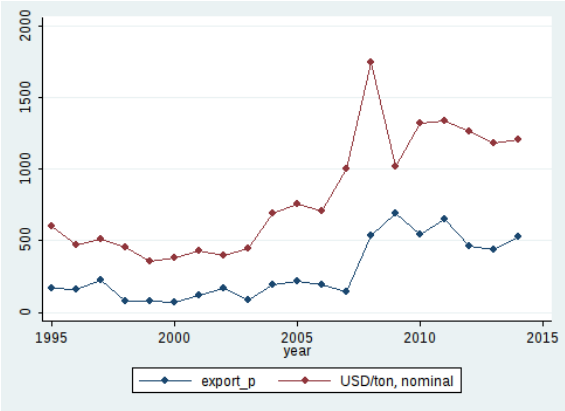
Figure 1: Comparison of USGS prices and US Export Prices (Red line is USGS provided price per ton, blue is US export price, in USD per ton)



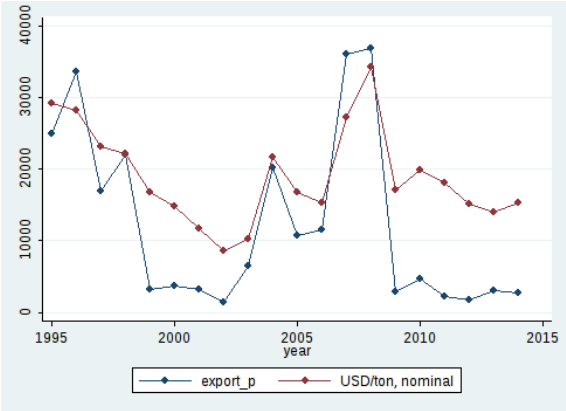
(a) Antimony



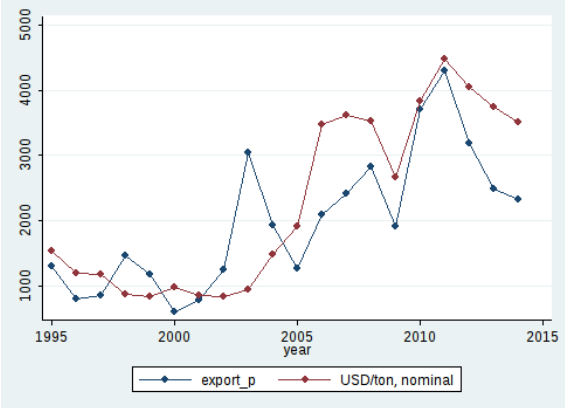
(b) Bismuth



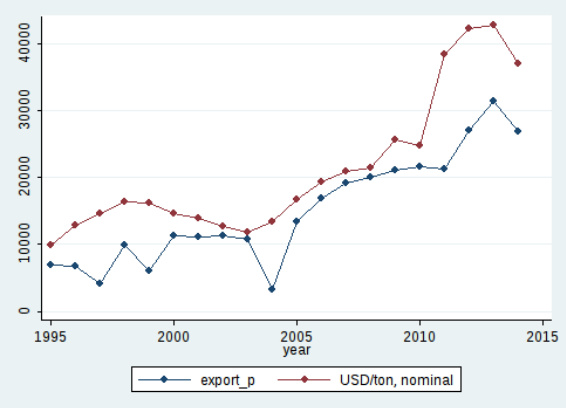
(c) Chromium



(d) Cobalt

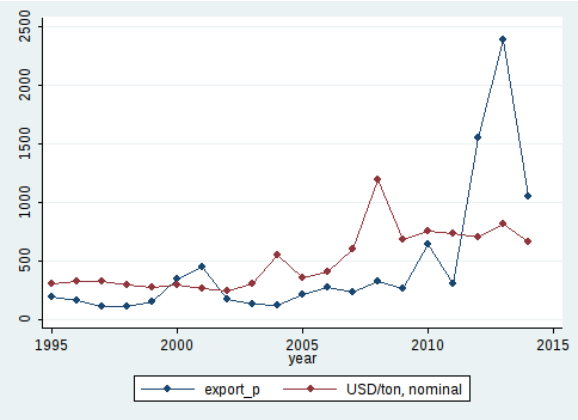


(e) Copper

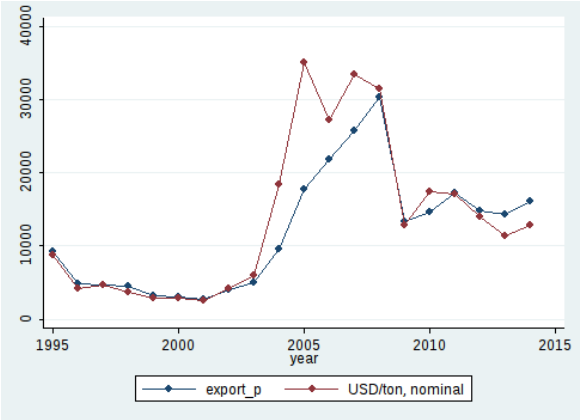


(f) Iodine

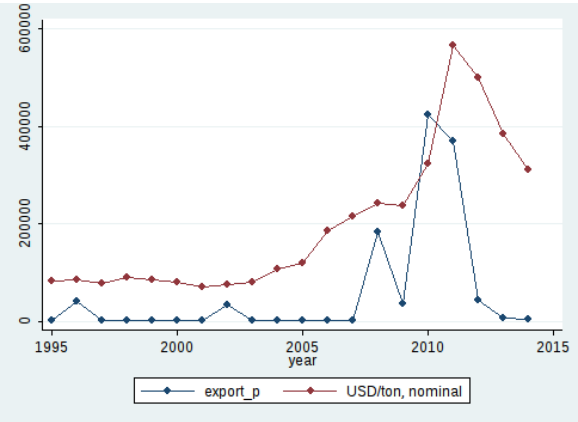
Figure 2: Comparison of USGS prices and US Export Prices (Red line is USGS provided price per ton, blue is US export price, in USD per ton)



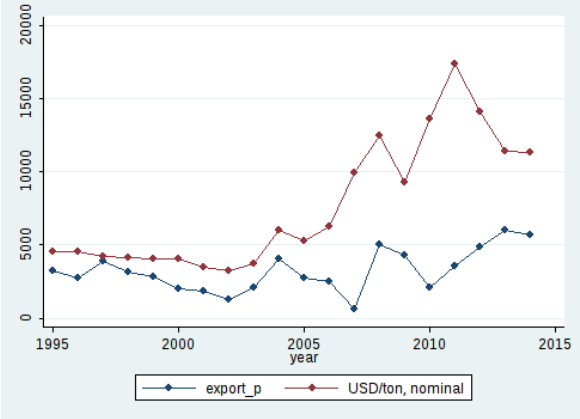
(a) Manganese



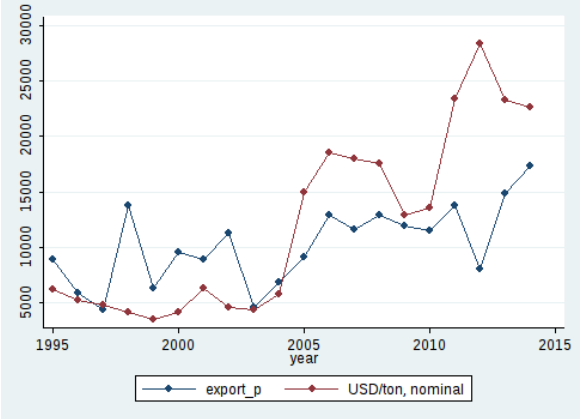
(b) Molybdenum



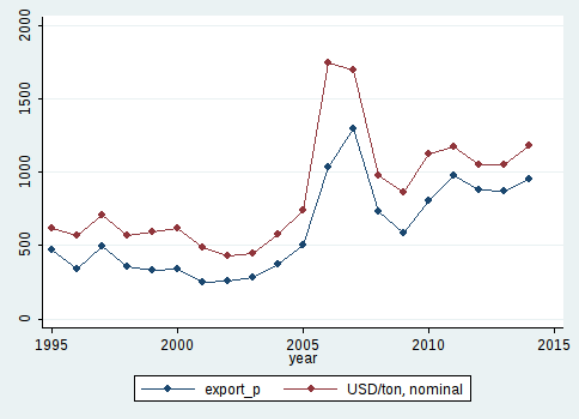
(c) Silver



(d) Tin



(e) Tungsten



(f) Zinc