Mauro Caselli, Jörg Mayer, Adrian Wood Revised January 2011

Explanatory Note on Dataset

for 'Has China de-industrialised other developing countries?'

Contents

1. Introduction and overview

This note explains how the dataset developed for this paper is organised and where all the data and more detailed explanations can be found. The final version of the paper used only parts of the dataset (subsets of countries, years and variables), but the entire dataset is being made available in the hope that it will be useful for other research. It is a compilation of data, some not available elsewhere, on the sectoral composition of output and trade, and on factor endowments, in many countries over a long period.

1.1 Files included in the dataset

Full Dataset.xls includes all the data, and the sources and descriptions of all variables, for 133 countries, annually between 1970 and 2005. These 133 countries were chosen on the basis of having populations above 1 million in 1990 and data on average years of schooling (though six – Botswana, Eritrea, Lesotho, Namibia, Puerto Rico and Yemen – which met both these criteria were omitted because they had no trade data). There are many gaps in the other series, especially in some regions (e.g. countries of the former Soviet Union) and for some variables (particularly sectoral output data and shares of the population with specific levels of schooling);

Panel Dataset.xls is a subset of the full dataset, including only data at 5-year intervals between 1970 and 2000, also with sources and definitions. In addition, it includes the dependent variables derived for the regressions in the paper, which are 3-year centred averages of logged output and export ratios (and corresponding 3-year-average output and export shares, though these were not used in the paper);

Unido Dataset.xls includes both the original data from UNIDO and the derived data included in the Full and Panel datasets (section 3 on 'Derivation of the output data' provides more information on their derivation and modification);

Imputed AVSY data.xls includes the calculations for the imputations to fill gaps in the average years of schooling data, using UNESCO data on literacy rates, as explained in section 2 of this note; of the 133 countries, no imputations were needed for 95, imputations for some but not all years for 19, and imputations for all years for 19.

1.2 How the data files are organised

The variables included in the Full and Panel datasets are defined, and their sources given, in the 'sources and definitions' sheets of the relevant workbooks. The units are specified at the top of the columns containing the variables. The columns are grouped into the following blocks.

Basic: country, country code, region and year.

Resources: land area (World Development Indicators), adult population (Population Division of the UN), average years of schooling of adult population (Barro and Lee, 2001, and Cohen and Soto, 2001, plus calculations described in section 2 of this note), and shares of the adult population with specific levels of schooling (Barro and Lee, 2001). Also included are trade/GDP ratios and PPP prices (Penn World Tables).[1]

Production: GDP (UNCTADstat online database); value added in agriculture, mining and manufacturing (UN National Accounts); and a division of UNIDO manufacturing value added data into four sectors (explained in section 3 of this note). Output shares of two major sectors: broad primary[2] and narrow manufactures, the latter subdivided between labour-intensive, skill-intensive, and electronics.

Exports and *Imports* (UNCTAD data)[3] include totals and disaggregations into sectoral categories compatible with the output data: narrow (unprocessed) primary, processed primary, labour-intensive narrow manufactures, skill-intensive narrow manufactures and electronics. Primary trade, both processed and unprocessed, is divided between mining and agriculture (and within agriculture between 'static' and 'dynamic' goods). The allocation of goods to these categories by SITC codes is described in section 4 of this note.

Dependent Variables are ratios derived from the production, export and import data in earlier blocks, and are included only in the Panel Dataset. Some related shares are also included in this section.

The production, export, import and dependent variable blocks all contain data using two alternative specifications of the output and trade variables. Specification 1 puts electronics in skill-intensive manufactures for developed countries (regions 1 and 2) but in labour-intensive manufactures for developing countries. Specification 2 puts

---

[1] These variables need to be interpreted cautiously. The trade ratio in column S is the trade variable in column T (half of exports + imports) divided by the GDP data in col Z, with both trade and GDP taken from the UNCTAD database. This trade ratio can be adjusted to put its denominator on to a PPP basis by multiplying it by the p variable in column X (from the Penn World Tables), which makes it smaller for most countries, especially poorer ones. However, these adjusted trade ratios are different from the openpppp ratios in columns U and V, which come directly from the Penn World Tables, for reasons that we have not been able to discover.

[2] Using the data provided, broad primary output can be split into processed primary and unprocessed primary (the latter being further divisible between agriculture and mining). See section 3 below.

[3] [3] Our data are similar to those in the UNCTADstat online database, but go back further (years prior to 1995 are not included in UNCTADstat) and are based on SITC Rev. 2 (whereas UNCTADstat is based on Rev. 3). We downloaded our data in 2006-7, prior to the creation of UNCTADstat (which uses better methods of estimating missing data).

electronics in skill-intensive manufactures for all countries.

2. Derivation of the education data

This section explains how gaps in the data on average years of schooling were filled, how workers were assigned to skill categories, and how gaps in schooling data for China and Germany were filled.

2.1 Average years of schooling imputations

The coverage of the Barro-Lee (2001) and Cohen-Soto (2001) data on average adult years of schooling was expanded using UNESCO data on literacy rates. This section describes the methods used to make the imputations and explains the decisions that had to be taken and the reasons for choosing a particular option. The results for the Barro-Lee data are shown in *Imputed AVSY data.xls*

The imputations are made for those observations for which there are data for literacy rates but not for average years of schooling. Thus, actual average years of schooling (AVSY) are used when available, while the imputed data are used in other cases.

Gaps in the literacy rate data for 1970-2000 are filled through linear interpolations, particularly for 1975, when data are available only for a few countries. For the period prior to 1970, too few literacy data are available to make the calculations feasible.

Regressions were run on those countries for which both sets of data are available in order to estimate the relationship between AVSY and literacy rates. The final set of imputations is based on countries with populations above 1 million in 1990, reported in the 'imputations, countries > 1 mill' sheet of *Imputed AVSY data.xls*, and is derived from the following regression:

$$\ln y_{it} = \alpha + \beta x_{it} + \gamma x_{it}^2 + \delta x_{it}^3,$$

where y represents average years of schooling taken from the Barro-Lee dataset, x represents literacy rates based on UNESCO data and the subscripts i and t stand for the country and the year, respectively. The regression includes power terms because it performs better than a linear equation in two respects: matching the actual Barro-Lee AVSY data and matching the data in the Cohen-Soto dataset.

The imputations based on this regression were preferred over those based on the following regressions:

$$y_{it} = \alpha + \beta x_{it} + \gamma x_{it}^2 + \delta x_{it}^3$$
$$y_{it} = \alpha + \beta x_{it} + \gamma x_{it}^2 + \delta x_{it}^3 + \eta z$$
$$\ln y_{it} = \alpha + \beta x_{it} + \gamma x_{it}^2 + \delta x_{it}^3 + \eta z,$$

where z represents year dummies. The reason why the log-linear specification was preferred to the linear one is that the latter predicts some negative values of AVSY for Burkina Faso and Chad, which cannot happen in the log specification. The reason

why the year dummies were not included in the final specification is that they lead to a higher variance of the error term.

For a few countries, the Barro-Lee data cover only from 1 to 4 of a maximum of 7 years. This presents an additional complication in the sense that if one were to use the actual Barro-Lee data for the available years and imputed data for the missing years, the time series of data for the country concerned would be internally inconsistent. This requires some form of adjustment of the imputed data for the missing years.

Two adjustment methods were explored, absolute and proportional. The absolute adjustment is based on differences, and the proportional adjustment on ratios, between actual and imputed values for years in which both are available. To fill gaps between two years for which both actual and imputed data are available, the difference or ratio is interpolated linearly. To extend beyond a year for which both actual and imputed data are available, the difference or ratio for the year concerned is held constant. The results of the absolute and proportional adjustments were similar, but the proportional one was used for the final calculations because the absolute method predicts negative values of AVSY in a few cases.

2.2 Education categories

Following Wood (1994), the labour force in each country is divided into three skill categories: labour with no education (unskilled workers), with basic education (low-skilled workers), and with substantial post-basic education and training (high-skilled workers). Workers with no education are generally unsuitable for employment in manufacturing, so people without any formal education (not even incomplete primary schooling) were always placed in the unskilled category.

The educational boundary lines around low-skilled and high-skilled workers are less obvious, especially because skilled workers are a mixture of those with only a basic education but much training (craftsmen), and those with substantially more education (professional and technical workers). Four alternative sets of boundaries were tried, of which specification 1a yielded the best statistical results. The acronyms refer to the number of adults for whom the given level of education is the highest attained:

TH = complete or incomplete higher education
CS = complete secondary education
IS = incomplete secondary education
TS = CS+IS = complete or incomplete secondary education
CP = complete primary education
IP = incomplete primary education
TP = CP+IP = complete or incomplete primary education

Specification 1:	High 1 = TH	Low 1 = TS+TP;
Specification 1a:	High 1a = TH	Low 1a = TS+CP
Specification 2:	High 2 = TH+CS	Low 2 = IS+TP
Specification 2a:	High 2a = TH+CS	Low 2a = IS+CP

In specifications 1 and 2, the unskilled category is simply workers with no education, while in specifications 1a and 2a, the unskilled category implicitly includes also those with incomplete primary education (in the dataset it is $u + b1 - b1a$ or $u + b2 - b2a$).

2.3 Estimation to fill gaps for China and Germany education categories

The breakdown of the labour force by schooling categories is not available in the Barro-Lee data set for China in 1970. This gap was filled in the following way.

The principle was to find average years of schooling for each category in 1975 for China that would generate the Barro-Lee average years of schooling of adult population (BL AVSY) as a weighted average of these individual averages, using the shares of the skill categories as weights, and then, assuming the same average years for each category in 1970, to find shares that would generate the BL AVSY for 1970 and also be consistent with the sizes and directions of later changes in these shares (with allowance for the Cultural Revolution 1966-76, which particularly affected university education).

The same principle was applied to fill gaps for Germany's education categories. In this case data for 1985 and 1995 were used to impute values for 1990.

3. Derivation of the output data

This section explains how output data for primary products, *P*, and labour-intensive manufactures, *M*, were derived by combining UN national accounts statistics with UNIDO industrial output statistics.

The primary category, *P* (called *BP* in our earlier work, e.g. Wood and Mayer, 2001), is close to categories 0-4 of the Standard International Trade Classification (SITC). It therefore includes processed primary products (*PP*), which the International Standard Industrial Classification (ISIC) treats as manufactures. More precisely,

P (broad or SITC primary) = *NP* (narrow or ISIC primary) + *PP* (processed primary);
NM (narrow or SITC manufacturing) = *BM* (broad or ISIC manufacturing) – *PP*.

In our trade data, *NM* (roughly SITC categories 5-9) is divided into labour-intensive manufactures (*NML*), skill-intensive manufactures (*NMH*) and electronics (*EL*), using the SITC categories specified in section 4 of this note, so that

$$BM = PP + NM = PP + NML + NMH + EL.$$

The UN national accounts and UNIDO data, however, use ISIC categories, for which it was necessary to specify a concordance with our SITC-based categories. Moreover, the UN national accounts data in principle cover the whole of the economy, but at a high level of aggregation, whereas the more disaggregated UNIDO data refer only to broad manufacturing (and exclude the output of small firms). The UN and UNIDO data thus also had to be combined, in the following way.

For each country and year in which they were available, the UNIDO data on broad manufacturing value added were allocated among our categories as follows, where the numbers refer to ISIC (Rev. 2) 3-digit groups:

PP: 311, 313, 314, 353, 354, 372
NML: 321, 322, 323, 324, 331, 332, 341, 342, 355, 356, 361, 362, 369, 371, 381, 390
NMH: 351, 352, 382, 384, 385
EL: 383

The resulting numbers were used to calculate four shares: $\psi_{PPBM} = PP/BM$; $\psi_{NMLBM} = NML/BM$; $\psi_{NMHBM} = NMH/BM$; and $\psi_{ELBM} = EL/BM$. These shares were then used to convert National Accounts sectoral output data for agriculture (*NPA*), mining (*NPM*) and ISIC manufacturing (BM_{NA}) into *P* and *NM*, as follows,

$P = NPA + NPM + (\psi_{PPBM} * BM_{NA})$
$NM_{NA} = (1 - \psi_{PPBM}) * BM_{NA}$

and to split up narrow manufacturing, as follows

$NML_{NA} = \psi_{NMLBM} * BM_{NA}$;
$NMH_{NA} = \psi_{NMHBM} * BM_{NA}$;
$EL_{NA} = \psi_{ELBM} * BM_{NA}$.

Our category *M* is defined in two alternative ways, as noted above. In specification 1, for developing countries *M* is $NML_{NA} + EL_{NA}$, and for developed countries is NML_{NA}; while in specification 2, *M* is NML_{NA} for all countries.

We encountered four problems in deriving the required output data from the UNIDO Industrial Statistics Database 2006:

1) For some countries and years, some data were aggregated across 3-digit groups. To disaggregate them, we applied the shares of the relevant groups in the aggregate in the previous three years for which disaggregated data were available. Countries for which there were no disaggregated data for any year were excluded.

2) Missing data for some groups and years. Such gaps were filled (i) by calculating an average value if the missing observation was between two available ones, or (ii) by calculating the percentage change for the three previous or following years, depending on whether the observations were missing at the end or at the beginning of a period. Countries for which these calculations could not be made were excluded.

3) Data for some countries had to be combined to obtain consistent time series. Data for West Germany (available up to 1993) were combined with data (available from 1998) for Germany (which includes the former East Germany). Data for Belgium and Luxembourg were combined to match available trade data. Data for Ethiopia were combined with data for Eritrea, using the dollar exchange rates in the UNIDO dataset to convert the data for Eritrea into Ethiopian Birr.

4) Part of the data for Burundi and for Trinidad and Tobago appears to be misreported. We thus assumed that for Burundi group 383 had been wrongly placed in group 390 from 1986 to 1991, and that for Trinidad and Tobago groups 382 and 383 had been wrongly aggregated into group 383 between 1982 and 1987.

4. SITC categories in each sector

Our trade data are based on the Standard International Trade Classification (SITC, Revision 2), and are at the 3-digit level. This section explains how SITC categories were assigned to the sectoral groupings used in this paper. The same groupings were used in Wood and Mayer (2001) and Mayer and Wood (2001), but the data were from a different source (and the allocation of SITC categories to the primary groupings was not quite the same: it is specified in Annex 1 of Wood and Mayer, 1998).

4.1 Division between manufactures and primary products

Narrow manufactures (**NM**) correspond to SITC categories 5–9, and broad primary products (**BP**) to SITC categories 0–4, except that the following items in SITC 5–9 are classified as BP rather than NM:

	SITC-Rev 2 categories
Radioactive and associated material	524
Pearls and precious stones	667
Non-ferrous metals	68
Zoo animals and pets	941
Non-monetary gold	971

4.2 Subdivision of manufactures

Narrow manufactures (NM) are subdivided into three categories: labour-intensive manufactures (NML), skill-intensive manufactures (NMH), and electronics (EL). The division between labour-intensive and skill-intensive items is based on a review of earlier studies that ranked manufacturing industries by their skilled/unskilled labour ratios or other measures of skill intensity, particularly the studies surveyed in Wood (1994, ch. 3) and OECD (1992). Electronics include computers and office equipment, and communication equipment and electrical machinery.

The SITC categories allocated to these sectoral groupings are as follows:

Labour-intensive manufactures (**NML**)	SITC-Rev 2 categories
Leather and rubber products	61–62
Wood and paper products	63–64
Textiles, clothing, travel goods and footwear	65, 83–85
Non-metallic mineral products, excl. precious stones	66 less 667
Iron and steel and metal products	67, 69
Furniture and plumbing equipment	81–82
Ships, bicycles and trains	78 (less 781-784), 79 (less 792)
Miscellaneous	89, 9 (less 941, 971)

Skill-intensive manufactures (**NMH**)	SITC-Rev 2 categories
Chemicals	5 (less 524)

Non-electrical machinery	71–74
Motor vehicles and aircraft	781–784, 792
Scientific instruments, watches and cameras	87, 88

Electronics (**EL**)

Computers and office equipment, and communication equipment and electrical machinery	75–77

4.3 Subdivision of primary products

Broad primary products (BP) are subdivided into the following six groups, to which the allocation of SITC 3-digit categories is set out in more detail further below.

PPM	processed minerals, metals and fuels
PPD	processed 'dynamic' agricultural products
PPS	processed 'static' agricultural products
NPM	unprocessed minerals, metals and fuels
NPD	unprocessed 'dynamic' agricultural products
NPS	unprocessed 'static' agricultural products

The distinction between 'dynamic' and 'static' agricultural export sectors is based on their income elasticities of demand (for further discussion, see Annex 1 of Wood and Mayer, 1998).

Other product groupings are simple aggregations of those defined above:

PP	processed primary products (= PPM + PPD + PPS)
NP	unprocessed ('narrow') primary (= NPM + NPD + NPS)
BM	broad manufactures (= NM + PP)
BPM	minerals, metals and fuels (= PPM + NPM)
BPA	agricultural products (= PPD + PPS + NPD + NPS)
PPA	processed agricultural products (= PPD + PPS)
NPA	unprocessed agricultural products (= NPD + NPS)

The following list shows which SITC 3-digit categories were allocated to each of the six primary product groups defined above (PPM, PPD, PPS, NPM, NPD, NPS).

Product Category	**SITC, Rev 2**	**Product**
I. <u>FOODSTUFFS and TOBACCO</u>		
LIVE ANIMALS	001	NPS
MEAT		
Fresh, chilled or frozen	011	NPD
Salted, in brine, dried or smoked	012	NPD
Prepared or preserved, nes; fish extracts	014	PPD

DAIRY PRODUCTS		
Milk and cream	022	NPD
Butter	023	PPD
Cheese and curd	024	PPD
EGGS	025	NPD
FISH and SHELL FISH		
Fish, fresh or frozen	034	NPD
Fish, dried, salted, smoked	035	NPD
Shell fish, fresh or frozen	036	NPD
Fish and shell fish, prepared or preserved	037	PPD
WHEAT, BARLEY, MAIZE and other CEREALS		
Unmilled cereals	041, 043–045	NPS
Wheat flour and meal	046	PPS
Other cereal meals and flours	047	PPS
Cereals etc preparations	048	PPS
RICE	042	NPS
VEGETABLES		
Vegetables, fresh, simply preserved	054	NPD
Vegetables, prepared or preserved	056	PPD
FRUITS and NUTS		
Fruits and nuts, fresh or dried	057	NPD
Fruits, preserved, prepared	058	PPD
SUGAR		
Sugar and honey	061	NPS
Sugar confectionery and preparations	062	PPS
COFFEE	071	NPS
COCOA		
Cocoa	072	NPS
Chocolate and chocolate products	073	PPS
TEA and MATE	074	NPS
SPICES	075	NPD
FEEDING STUFF FOR ANIMALS	081	PPS
MARGARINE and SHORTENING	091	PPS
EDIBLE PRODUCTS AND PREPARATIONS NES	098	PPS

NON-ALCOHOLIC BEVERAGES	111	PPS
ALCOHOLIC BEVERAGES	112	PPS
TOBACCO		
Tobacco, unmanufactured	121	NPS
Tobacco, manufactured	122	PPS
OILSEEDS		
Seeds for soft fixed oils	222	NPS
Seeds for other fixed oils	223	NPS

II. AGRICULTURAL RAW MATERIALS

LEATHER		
Hides and skins, excluding fur skins	211	PPS
Fur skins	212	PPS
RUBBER		
Natural rubber, gums	232	NPS
Rubber, synthetic, reclaimed	233	NPS
WOOD		
Cork	244	NPS
Fuelwood	245	NPS
Other wood, rough	247	NPS
Sawnwood and sleepers	248	PPS
PAPER, PAPERBOARD, PAPER PULP		
Pulpwood	246	PPS
Pulp and waste paper	251	PPS
SILK	261	NPS
FIBRES		
Cotton	263	NPS
Jute	264	NPS
Vegetable textile fibres (other than cotton and jute)	265	NPS
Synthetic fibres for spinning	266	NPS
Other man-made fibres	267	NPS
Waste of textile fibres	269	NPS
WOOL AND OTHER FINE ANIMAL HAIR	268	NPS
CRUDE ANIMAL MATERIALS	291	NPS
CRUDE VEGETABLE MATERIALS	292	NPS
ANIMAL OILS and FATS		
Unprocessed	411	PPD

Processed	431	PPS
VEGETABLE OILS		
Fixed, soft	423	PPD
Other fixed	424	PPD
III. ORES, MINERALS and METALS		
FERTILIZERS	271	NPM
CRUDE MINERALS		
Stone, sand and gravel	273	NPM
Sulphur	274	NPM
Natural abrasives (including industrial diamonds)	277	NPM
Other crude minerals	278	NPM
IRON		
Iron ore and concentrates	281	NPM
Iron and steel scrap	282	NPM
URANIUM and THORIUM		
Ores and concentrates	286	NPM
Radioactive and associated materials	524	PPM
Uranium and thorium unwrought or wrought and articles thereof	688	PPM
BASE METALS		
Ore and concentrates	287	NPM
Non-ferrous metal scrap	288	NPM
Copper	682	PPM
Nickel	683	PPM
Aluminium	684	PPM
Lead	685	PPM
Zinc	686	PPM
Tin	687	PPM
Non-ferrous base metals	689	PPM
SILVER, PLATINUM and GOLD		
Ores and concentrates of precious metals	289	NPM
Silver and platinum unwrought, unworked or semi-manufactured	681	PPM
Gold, non-monetary	971	NPM
IV. NON-METALLIC MINERAL MANUFACTURES		
Pearls, precious and semi-precious stones	667	NPM

V. FUELS AND ELECTRIC ENERGY		
COAL	322	NPM
COKE and BRIQUETTES	323	NPM
PETROLEUM		
Crude petroleum	333	NPM
Petroleum products, refined	334	PPM
Residual petroleum products	335	PPM
GAS Natural and Manufactured	341	NPM
ELECTRIC ENERGY	351	PPM
VI. OTHER		
Zoo animals, pets	941	NPS

5. References

Barro, R., and J.-W. Lee (2001). International data on educational attainment: updates and implications. *Oxford Economic Papers*. 53(3): 541–563.

Cohen, D., and M. Soto (2007). Growth and human capital: good data, good results. *Journal of Economic Growth*. 12(1): 51–76.

Heston, A., R. Summers, and B. Aten (2006). Penn World Table Version 6.2. Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania.

Mayer, J. and A. Wood (2001). South Asia's export structure in a comparative perspective. *Oxford Development Studies*, 29(1): 5-29.

Organisation for Economic Co-operation and Development (1992). *Industrial Policy in OECD Countries: Annual Review 1992* (Paris: OECD).

Wood, A. (1994). *North-South Trade, Employment and Inequality*. Oxford: Clarendon Press.

Wood, A., and J. Mayer (1998, 2001). Africa's export structure in a comparative perspective. The 1998 working paper version is Study No. 4 of *African development in a comparative perspective* Geneva: UNCTAD, and is available on request from the authors. The 2001 version is in the *Cambridge Journal of Economics*, 25(3): 369-94.