A COMPOSITIONAL STATISTICAL ANALYSIS OF CAPITAL STOCK

Juan M. Larrosa
CONICET – Universidad Nacional del Sur – Universidad Empresarial Siglo 21 (Argentina);
jlarrosa@criba.edu.ar

Abstract

Most of economic literature has presented its analysis under the assumption of homogeneous capital stock. However, capital composition differs across countries. What has been the pattern of capital composition associated with World economies? We make an exploratory statistical analysis based on compositional data transformed by Aitchinson logratio transformations and we use tools for visualizing and measuring statistical estimators of association among the components. The goal is to detect distinctive patterns in the composition. As initial findings could be cited that:

1. Sectorial components behaved in a correlated way, building industries on one side and , in a less clear view, equipment industries on the other.
2. Full sample estimation shows a negative correlation between durable goods component and other buildings component and between transportation and building industries components.
3. Countries with zeros in some components are mainly low income countries at the bottom of the income category and behaved in a extreme way distorting main results observed in the full sample.
4. After removing these extreme cases, conclusions seem not very sensitive to the presence of another isolated cases.

1. Introduction

While physical capital stock represents a crucial factor in the economic process, less is known about the joint behavior of capital components. This paper tries to show first results about how the composition of capital has performed during the 1965-1990 period for a heterogeneous sample of countries. We used statistical tools for visualizing patterns in the data sample as well as recent economic evidence to show some possible explanations.

Given that we are asking about capital components, we should use data that reflects its composition and variability. We used compositional data that consists of positive valued vectors summing to a unit (hundred per cent). Examples of this kind of data in Economics are many, including household budget shares, aggregate output, stockholder’s portfolio composition, etc. Several issues condemn this type of data for using typical statistical inference methods. It follows that some transformation, if it exists, has to be applied before analysis. Fortunately in our case it exists, and allows for the use of almost full multivariate analysis procedures. Our goal is to find patterns in the capital per worker composition looking for answers about how these components have performed. This behavior should be interpreted as the struggle among economic sectors for capital allocation. Due to the small number of available components we only found in most samples analyzed a common behavior of sectorial components identified as equipment and building sectors components. In any case, behavior seems to be highly sensitive to the presence of extreme cases. The process of identification of extreme cases is sequential. We begin by analyzing full sample data and we follows with subsamples defined by income level categories. After detecting extreme cases we redo former analysis excluding these outliers and we arrive to the final conclusions.

The paper is organized as follows. Section 2 summarizes recent literature on countries’ physical capital investment behavior. Section 3 describes the statistical theory and definitions that supports the analysis. Section 4 presents the results of sample and subsample analysis and section 5 ends with preliminary conclusions and discussion.

2. Literature on physical capital patterns

Several works have emphasized the importance of specific capital investment as requirements for growth. Since De Long and Summers (1990) shaded light to the roll of equipment investment in the growth process for a sample of countries during the period 1960-1985, many other research works supported this finding in the broad sense (for example, Temple and Voth, 1998.) At the same time, Jones (1994) investigated how affected is growth by distortions in capital relative price. Working with some of the same variables of this paper, Jones found that higher relative prices of capital (through taxes or tariffs on importing) resent growth. Explicitly, he found negative correlation between all capital subaggregate components relative prices and annual growth rate per capita. In a more theoretical framework, Jovanic and Rob (1997) used a modified Solow growth scheme for modeling the observed fact that machinery is more expensive in less developed countries. They replicated reasonably well real data and the conclusion of their work points out the relative shortage in machinery participation in less developed countries something also observed in this work. Seitz (1995), using German regional data, found that public capital stock provision was a sensible input in the private sector production function and that public capital acted as complementary to private capital. Externalities appear mainly through transportation cost reduction. In another work, Seitz (2000) found that urban infrastructure affects city competitiveness by reducing costs in local firms because agglomeration externalities. Devarajan, Swaroop, and Zou (1996) divided public expenditure between current expenditure and capital investment, they defined them as unproductive and productive capital respectively. They found empirically that long run growth is positively correlated with public capital investment and negatively related with public current expenditure. In a labored theoretical paper, Turnovsky and Fisher (1995) developed a framework for the analysis of expenditure composition. They obtained a model that relates public expenditure (government consumption and public investment in infrastructure) with macroeconomic performance in an intertemporal optimization model. Interestingly the authors allow in the model to government consumption to be complementary to private consumption and work effort. They found that infrastructure investment promoted by government could provoke a negative effect through a contraction in the short run for the displacement of resources from public to private sector but improves the welfare in the long run for the better future conditions for economic activity. Finally, the most comprehensive research into particular components of capital stock of the economies could be found in a research paper series supported by the World Bank that will be following summarized.

Canning (2000) develops a panel data production function estimation that includes as infrastructure variables: miles of roads, electricity generating capacity, and telephones per workers. He found that only telephones per worker is statistically significant in the sample, suggesting that this variable generates more externalities in the economy than the first two. Ingram and Liu (1997) estimated the influence of economic variables in a wide range of equipment and transportation variables in a heterogeneous sample of countries and cities. Their work shed light on the pros and against of high level of motorization in big cities and the externality that this provokes in land prices, congestion, and pollution. As they recalled in another related paper (Ingram and Liu, 1999) in the past 15 years the World stock of vehicles grew up in about 60%, because of lower production costs and a higher relative income in less developed countries. This way it could be expected a significant participation of transportation capital in the total stock of capital (or at least an increase in recent years). Again, the question remains of whether this increment has been done by taken participation of another class of capital. Randolph, Bogetic, and Heffley (1996) found a set of variables that correlates positively with investment in infrastructure related to transportation and communication sector. These variables are the urbanization level, foreign sector size, population density, and funding mechanism, among others.

A crucial feature related to infrastructure investment is how these projects are funding and financing. Klingebiel and Ruster (2000) summarize that most governments induce private sector to invest in infrastructure through soft lending, guarantees, and grants with a wide variety of results. This inducement process has had very different results depending on the institutional framework implemented and the specific financed project, but this remarks how infrastructure market is an active one, not only wrapped around the government hand. But government-funded investment has a crucial roll in this aspect. Reinikka and Svensson (1999) studied the cases of less developed countries where in some cases they assured that government investment in infrastructure is even more important than macroeconomic indicators for the private sector investment decision process. Infrastructure provides through cost reductions and linkages positive externalities to economy as a whole.

At the same time, the building sector shows itself as a highly expansive one in whether developed and undeveloped countries. Housing is upraising in developed countries because people are moving from downtown to suburbia. This observed behavior is robust to different kinds of shocks like those studied by Glaeser and Gyourko (2001) for the American case. New construction is enhanced by relative lower land prices and lower mortgage rates in developed countries. In the other hand, in less developed countries housing represents a substantial part of the capital stock because their less industrialized profile.

As suggested earlier, physical capital components seems to be markedly complementaries. The building of a dam requires not only of concrete and rolling stones but also of road infrastructure and housing for the workers. Canning and Bennathan (2001) studied the social rate of return of generating electricity capacity and paved roads projects and showed that both kinds of projects reflects higher than average rates of returns when considered simultaneously. In isolation, both kinds of projects reflect lower than social rates of return. That's because when they considered investments' potential benefits against its construction costs, complementarities emerge in a crossed way. This supports the idea of considering a mix of capital components when analyzing infrastructure investment, a key issue in the interpretation of the present work that we'll consider as the complementarity approach.

Another kind of physical capital is inventories. Guasch and Kogan (2001) survey the inventories statistics of a sample of countries and found that less developed countries have three times more inventories stocks than developed countries. The problem associated with keeping high inventories is usually lack of efficiency in the industry structure, transforming this inefficiency into tangible results through lower benefits (lost transactions, delays in deliveries, high amount of immobilized capital). Again, the low rate of investment in new depots or warehouses and the small market size does not help much in solving the problem in developing countries. They found that inventories levels are correlated negatively with GDP per capita and a dummy variable that counts for infrastructure quality.

Table 1 concisely reports main findings of the literature review and focuses in the main variables related to physical components analyzed by each research paper.

Table 1. Summary of empirical references

Author/s	**Capital Component**	**Results (type of data or analysis)**
De Long and Summers (1990)	Equipment and machinery investment	Positive correlation between growth rate and equipment and machinery investment (country data).
Temple and Voth (1998)	Equipment and machinery	Positive correlation between growth rate and equipment and machinery investment (country data)
Jovanovic and Rob (1997)	Equipment and machinery	Machinery is relative more expensive in less developed countries (country data)
Seitz (1995)	Physical capital	Presence of complementarity among capital components (regional data)
Seitz (2000)	Physical capital	Infrastructure investment, among other variables, affects city productivity (urban data).
Devarajan, Swaroop, and Zou (1996)	Public capital	Negative correlation between public current expenditure and long run rate of growth and positive correlation between public capital investment and long run rate of growth
Jones (1994)	Physical capital and components relative price	Negative correlation between capital component relative prices and growth (country data)
Canning (2000)	Non-residential construction and transportation equipment	A variable telephone per worker is statistically significant in explaining countries' aggregate output (country data).

Ingram and Liu (1997)	Durable goods and transportation equipment	Geographic and economic (country and urban) variables significantly correlated with motorization and transportation variables.
Ingram and Liu (1998)	Durable goods and transportation equipment	Environment and economic (country and urban) variables significantly correlated with motorization and transportation variables.
Randolph, Bogetic, and Heffley (1996)	Transportation equipment	Social, economic and institutional variables significantly correlated with public investment in transportation infrastructure (country data)
Klingebiel and Ruster (2000)	Infrastructure investment	Importance of private sector participation in infrastructure provision (case studies)
Reinikka and Svensson (1999)	Infrastructure investment	Importance of government infrastructure investment in private sector investment expectations (firm data)
Glaeser and Gyourko (2001)	Residential building	Several economic, social, and infrastructure variables explained significantly housing rates (urban data)
Canning and Bennathan (2001)	Non-residential construction	Importance of considering mix capital components in infrastructure analysis –for including complementarities and externalities effects (country data).
Guasch and Kogan (2001)	Equipment investment (inventories)	Negative correlation between inventories level and GDP per capita and infrastructure quality dummy (country data)

An interesting question that remains unanswered is the potential displacement of one class of capital by another during the economic process. What component has displaced equipment investment in high developed countries according to De Long and Summers (1990)? How about the increasing participation of housing as revealed by Glaeser and Gyourko (2001)? How are complementarities present in capital composition as mentioned by Canning and Bennathan? We will see that some clues for these questions could be obtained by using capital compositional data and specific statistical techniques and procedures.

3. Model and statistical techniques

Statistical data used in this investigation are compositional data. Compositional data refers to proportions of a whole and because of that are subject to the constraint that the sum of its components is unit or a constant. This restriction does not allow for a immediate interpretation of the covariance structure due to the presence of spurious correlation. This was unnoticed or not properly treated for long time by academic research across several disciplines. For instance, Brandt, Monroe, and Williams (1999) described the procedures commonly utilized by political scientists for avoiding this restriction: (1) ignoring the compositional nature of the data, for example, by using independent equations for each component, (2) ignoring all but one component, for example, any model of unemployment or political party vote share, or (3) converting a multipart composition into a two-part subcomposition and then employing (2). They remarked, first, that all of these approaches ignore the deterministic structure of the correlation among components caused by the sum constraint; second, all approaches ignore the boundedness of the data and third, the subcompositional approach can mask (or create) substantively important variability in the data.

The problem related to the difficulties for understanding the ‘obscurity’ of the covariance structure of a compositional set was first noticed by Pearson (1897). Aitchinson (1986) developed the transformations required for dealing with this problem and many others related to this particular kind of data[1]. Those developments have led to the realization that so-called standard multivariate analysis designed for unconstrained multivariate data is entirely inappropriate for the statistical analysis of compositional data: product-moment correlation of raw components is a meaningless descriptive and analytical tool in the study of compositional variability. As Aitchinson (1997) remarks: since there is a one-to-one correspondence

[1] Barceló-Vidal, Martín-Fernández, and Pawlowsky-Glahn (2001) formalized and stylized this framework.

between a composition and a complete set of ratios or logratios obtained from them, information remains the same in the process of transformation and these transformations possess some properties that are critical for compositional analysis: scale invariance, subcompositional coherence, meaningful groups of operations of change such as perturbation and power, meaningful measures of distance between compositions, among others. This section resumes the required concepts for understanding the findings of this work. We begin by defining what is compositional data.

Definition 1. Compositional data $x = (x_1, x_2, \ldots, x_D)'$ with D parts, is a vector with strictly positive components, so the sum of all of the components equal a constant *k*. The sampling space is the simplex defined as $S^D = \left\{(x_1, x_2, \ldots, x_D)' \; x_j > 0, j = 1, 2, \ldots D, x_1 + x_2 + \cdots + x_D = k\right\}$.

We can always obtain compositional data on S^D if we have an initial nonnegative components vector. We only require to divide each component by the sum of all components. Then we define:

Definition 2. The closure operator C is a transformation mapping each vector $w = (w_1, w_2, \ldots, w_D)'$ of R_+^D to its corresponding associated compositional data $C\,(w) = kw / (w_1 + w_2 + \cdots + w_D)$ of S^D, with k being the closure constant.

An important element of the analysis is the sample center or baricenter: Its definition is:

> **Definition 2.1**. The center or baricenter of a compositional data sample of size *N* is the geometric mean closure defined by $g_m = C\,(g_1, g_2, \ldots, g_D)$, where $g_i = \left(\prod_{n=1}^{N} x_i\right)^{1/N}, i = 1, 2, \ldots, D$.

In some cases it could be interested to reduce the dimensionality of the components by adding together a subsample of them. This procedure should be supported by theory or a requirement of the investigation under study.

Definition 3. Let *S* be a subset of 1,2,…, *D* of a compositional data $x \in S^D$ and being x_S a subvector formed by the corresponding parts of *x*, then $s = C\,(x_s)$ is called the subcomposition of the *S* parts of *x*.

In some other cases it is relevant for the investigation to focus the analysis in smaller number of components. We can use the closure operator on the sample of components and make the analysis as it were a composition in itself.

Definition 4. (Aitchinson, 1986, p. 37) If the parts of a *D*-parts composition are separated into *C* (≤ *D*) mutually exclusive and exhaustive subsets and the components within each subset are added together, the resulting *C*-part composition is termed *amalgamation*.

Another important tool for analyzing a compositional data set is the perturbation operator:

Definition 5. (Aitchinson, 1986, p. 42-43) Perturbation of one composition *x* by another composition *y* refers to the operation $x, y \in S^D \Rightarrow x \circ y = C\,(x_1 y_1, x_2 y_2, \ldots, x_D y_D) \in S^D$, which is termed a *perturbation* with the original composition *x* being operated on by the *perturbing* vector *y* to form a *perturbed* composition $x \circ y$.

Finally, the two main transformation we will apply to raw compositional vectors for its analysis: additive logratio transformation and centered logratio transformation.

Definition 6. Centered logratio transformation (*clr*) is a bijective application between $x \in S^D$ to $z \in \mathrm{R}^D$ defined by

$$clr(x) = \left(\ln \frac{x_1}{g(x)}, \ln \frac{x_2}{g(x)}, \ldots, \ln \frac{x_D}{g(x)} \right) = (z_1, z_2, \ldots, z_D),$$

with $g(x) = \left(\prod_{i=1}^{D} x_i \right)^{1/D}$ as the geometric mean of the composition. The inverse of the transformation in this case is $clr^{-1}(z) = \mathcal{C}\left(\exp(z_1), \exp(z_2), \ldots, \exp(z_D) \right) = (x_1, x_2, \ldots, x_D)$.

Notice that in *clr* transformation, geometric mean is estimated by using data matrix rows (observations) while in the definition of the center of observations set (ternary diagram center), geometric mean is calculated by columns (variables).

Vectors with up to three components can be easily visualized through a ternary diagram. This is a powerful tool for observing if data reflect some recognizable pattern. If this pattern exists, then a compositional straight line could pass through data in a way that captures most of observed points[2].

Throughout the paper we will widely use principal components analysis (PCA) based calculated using covariance matrix and biplots. Aitchinson and Greenacre (2000) extensively utilized these techniques and their paper represent an excellent review of the use of biplots and PCA for compositional data analysis. As another research papers that used compositional data and statistical tools could be mentioned Billheimer, Guttorp, and Fagan (1998) who modeled state-space models applied to Biology and Brehms, Gates, and Gomez (1998) using Dirichlet distributions in public administration studies.

4. Data structure and analysis

We begin this section by defining the relevant variables for this work. Data were extracted from Penn World Table 5.6 and correspond to KDUR, KOTHR, KNRES, KRES, and KTRAN series for the 1965-1990 time period. A brief description of these is published in Table 2. Series were selected only if they had full data series over the time period, and countries with zeros in any series were included only after applied the rounded zero replacement strategy proposed by Martín-Fernández, Barceló-Vidal, and Pawlowsky-Glahn (2000) and also suggested by Fry , Fry, and McLaren (2000).

Zeroes in a component usually are explained twofold: first, the variable is not really zero but because of lack of adequate measurement tools or techniques is often impossible or too expensive to obtain any meaningful or computable value for the variable so it is rounded as zero (these are called *rounded* zeroes). Second, the variable really takes zero value in some cases (these are the *essential* zeroes). In the second case we can't modify the value because we could alter the original real data, which probably belongs to a different population that of the one under study. In the second case it is justified to impute a 'small' value in order to process data by the logratio transformation.

Table 2. Code and description of variables

Index	Code	Description
1	**KDUR**	Percentage of capital per worker allocated in durable production assets (machinery and equipment).
2	**KOTHR**	Percentage of capital per worker allocated in other buildings.
3	**KNRES**	Percentage of capital per worker allocated in non-residential building.
4	**KRES**	Percentage of capital per worker allocated in residential building.
5	**KTRAN**	Percentage of capital per worker allocated in transportation equipment.

Source: Penn World Table 5.6

[2] Straight line is not a line as we could imagine for a two dimensional graph. Instead into the ternary diagram it seems more like a soft curve crossing for one side to the other.

Series were presented initially as percentages of the capital stock per worker in 1985 international prices. This fact made that total sum of components was different from unit in different periods. We proceeded by bounding the composition y closing each compositional vector year by year. This way we've got, for each year, the participation of each compositional vector in the hundred percent of each economy's capital stock per worker. Then we calculated the geometric mean of each vector for all the analysis period and closed it again because geometric means of variables were different than the total explanation. This way we obtain the average participation of each compositional vector for the time span of the sample.

World Bank (2000) defined subpopulations in terms of countries' level of income. Categories are: low income, lower middle income, upper middle income, high income (OECD countries), and high income (non-OECD countries). Strata are unequally covered due to our data availability constraint, existing 9 low income countries, 12 lower middle income countries, 9 upper middle income countries, 21 high income countries from OECD and 4 high income countries that are not OECD members. In Appendix raw data used in this work is published jointly with the country list and income category association.

Other indicators for clustering could be geographic indicators. Henderson, Shalisi, and Venables (2002) explain how spatial determinants affect economic outcomes in a wide variety of economic fields of study (urban economics, international commerce, and, specially for this investigation, international uneven distribution of production) by decisive agglomeration and network effects that affect relative prices and economic incentives. We could also use clustering techniques for identifying statistical subpopulations as suggested by Martín-Fernández, Barceló-Vidal, and Pawlowsky-Glahn (1998) but for the sake of clarity we considered this highly used classification.

Because of a possible small-sample-bias problem, we decided going to work with two main subsamples: high and low income. This way we deal with samples of reasonably size. High income category includes high income OECD countries and high income non-OECD countries and low income category includes low income, lower middle income, and upper middle income countries.

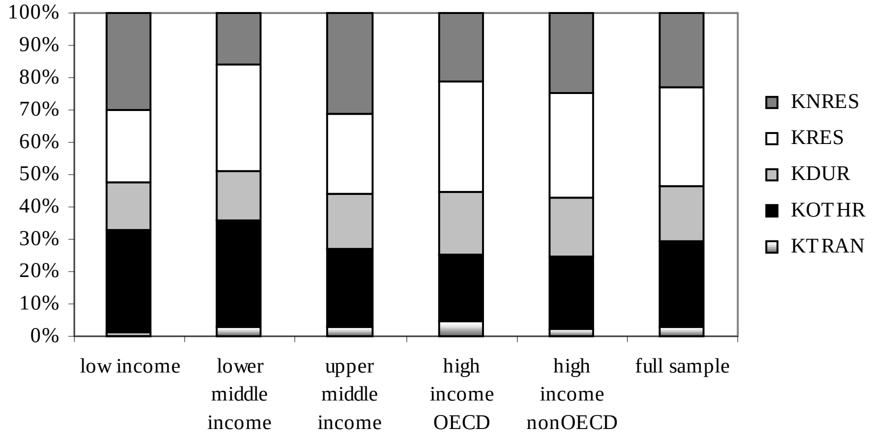
Once we obtained the final raw data block, we proceed to transform them with the centered logratio transformation *clr*. This imply that we should apply Definition 6:

$$clr(x) = \left(\ln \frac{x_1}{g(x)}, \ln \frac{x_2}{g(x)}, \ln \frac{x_3}{g(x)}, \ln \frac{x_4}{g(x)}, \ln \frac{x_5}{g(x)} \right)$$

with $g(x) = \left(\prod_{i=1}^{5} x_i \right)^{1/5}$ and 1,…,5 represents the index for the components in Table 2. Given that this transformation preserves the distance among data it becomes more useful for multivariate statistical analysis.

Full sample raw data descriptive statistics is published in Table 5 and Table 6 in the Appendix at the end of this paper. As we can see KTRAN is the most volatile variable, while KRES is the more stable compositional variable over the full sample. Figure 1 shows stacked bars for the full sample and all subsamples data of the five components. We can appreciate the differences between subsamples and full sample average. Transportation equipment is almost null in low income countries and got its highest average participation in the high income countries affiliated to OECD. Lowest income countries stand out by having a high share of their capital invested in other buildings and non residential construction. Lower middle income and high income countries distinguished themselves by showing a relative high part of their capital invested in residential building and an increasing share of transportation equipment relative to full sample average. At the same time, they showed a decreasing participation of capital allocated in other buildings.

Figure 1. Comparative raw data for income categories (sample average)



Data were tested for validation purposes. Aitchinson (1986, p. 143-148) proposed three tests for additive lognormal distribution detection: marginal test, bivariate angle test, and radius distribution test. The calculated values for the three tests are published in Table 3.

Table 3. Values of test statistics for logistic normality of data

			Anderson-Darling	Cramer –von Mises	Watson
Marginal	*i*				
	1		5.5748*	0.9004*	0.7704*
	2		0.7553****	0.1257****	0.1066****
	3		0.4902 +	0.0452 +	0.0368 +
	4		0.8364 ***	0.1186 ****	0.0925 +
	5		0.8897 ***	0.1345 ***	0.122 ***
Bivariate	*i*	*j*			
	1	2	1.7787 +	0.298 +	0.2713 *
	1	3	1.4087 +	0.2216 +	0.1516 +
	1	4	1.2523 +	0.1905 +	0.1599 ****
	1	5	0.8762 +	0.1317 +	0.1312 +
	2	3	0.6022 +	0.0869 +	0.074 +
	2	4	0.4522 +	0.0647 +	0.0645 +
	2	5	0.6035 +	0.08 +	0.0509 +
	3	4	1.1871 +	0.1677 +	0.0501 +
	3	5	0.6021 +	0.0966 +	0.0493 +
	4	5	0.467 +	0.063 +	0.0673 +
Radius			3.3987 **	0.3523 ****	0.1589 ****

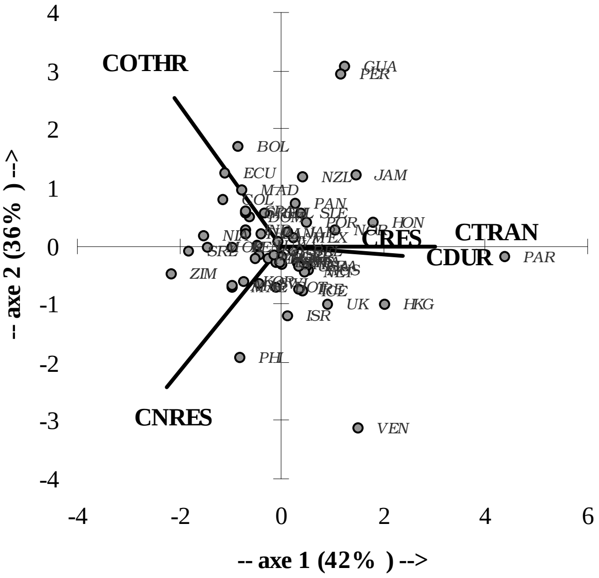
References
$p < .01$ * , $p < .025$ ** , $p < .05$ ***, $p < .1$****, $p > .1$ +

Tests weakly support the presence of logistic normality in the sample. The marginal and radius tests did not reject the logistic normality but bivariate test did. In fact, bivariate angle test shows significant departure from log normality. This way we can work on data that show some properties of logistic normality but these are not fully supported for the tests.

4.1 Full Sample Analysis

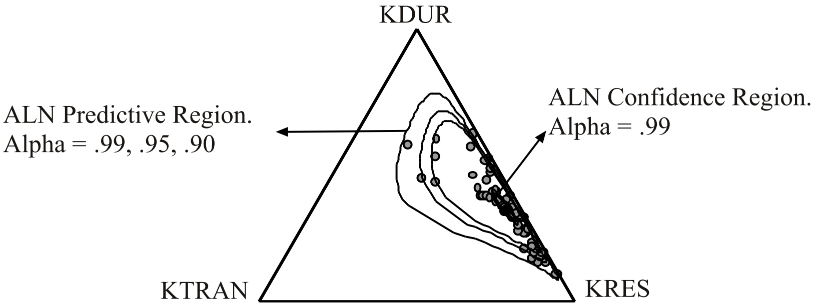
Clr-transformed data allow to full utilization of multivariate tests (transformed variables are denoted with a C instead of K prefix). PCA using the covariance matrix was calculated on the five compositional vectors and the biplot is showed in Figure 2 (total explained variability is between parenthesis). There, it can be checked out the magnitude and sign of the relationship illustrated in Figure 2.

Figure 2. Biplot on the first two principal components (78%) – Full sample



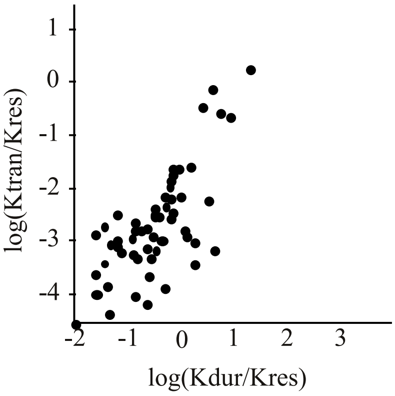
Almost coincident vertices are observed in CNRES, CDUR and CTRAN which behaves with scarce correlation with CNRES and COTHR. The first three variables seems to be, at different degrees, moving in the same direction and uncorrelated with the others two. Given the definition of the capital components, for the full sample it seems that housing, equipment and machinery production and transportation equipment behaves similarly, following increasing or decreasing participation in the capital stock during the economic process. The collinearity among these components could be better discerned by observing in the ternary diagram the data dispersion jointly with the corresponding additive log normal predictive regions (Fig. 3). The predictive regions at 99 per cent of significance level gather accurately all the data. For visualizations purposes data could be centered but in this case it won't be publish because the following development of the research does not require it.

Figure 3. Ternary diagram of subcomposition and additive lognornal predictive and confidence regions



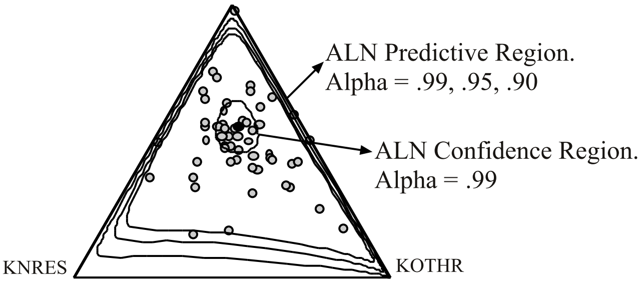
Another way of looking at this relationship is by plotting the log ratios among the variables and observing the clear linear relationship that results (Fig. 4).

Figure 4. Linear relationship among components KTRAN, KRES, and KDUR



Back to the five components analysis we must take into account the high heterogeneity of the sample. For this to be observed we amalgamated the three highly correlated components into one. Figure 5 shows in a ternary diagram the amalgamation KRDT = KRES + KDUR + KTRAN plotted with the other two components.

Figure 5. **Amalgamated data on the simplex (Raw Data - Full Sample)**

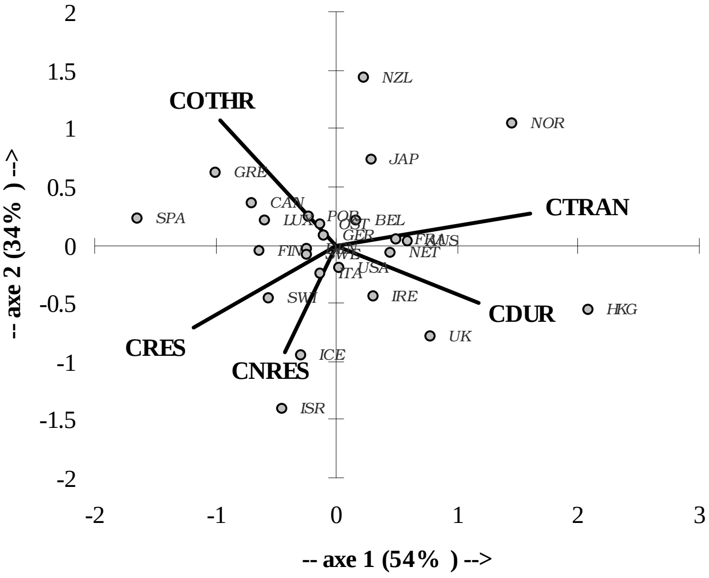


This observed relationship tell us about some possible joint behavior of two sector related to equipment manufacturing (KDUR and KTRAN) and one related to building sector (KRES). As mentioned, the ternary diagram showed in Figure 5 exhibits great data dispersion which could support the idea of the potential existence of different populations into the sample (total variance of 4.3695). The underlying heterogeneity of countries could be the reason of this variability. We will try to reduce it by clustering the sample.

4.2 Subsample Analysis

Now we follow the same procedure applied in Section 4.1 to the income-level-based clustering used by the World Bank. As mentioned earlier, Figure 5 showed the potential existence of different populations into the sample. We begin with high income sample by calculating the first two principal components (Fig. 6). Now the estimation shows collinear behavior between CRES and CNRES, in one hand, and CTRAN and CDUR, in the other hand. CRES and CTRAN have negative correlation (*r* = -0.673262) and CNRES and CDUR display scarce correlation (*r* = 0.06344) according to the displayed orthogonality.

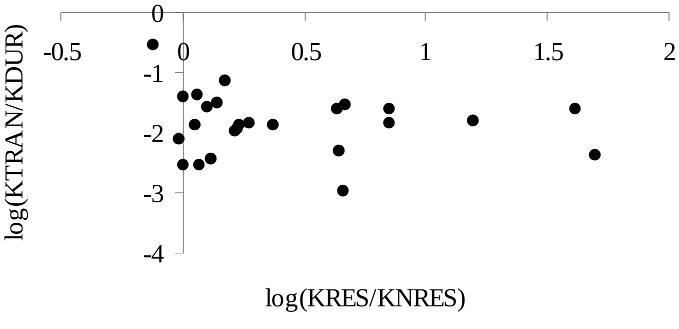
Figure 6. Biplot on the first two principal components (88%) – High income sample



These four, two *vis á vis*, show low correlation between them (*r* = -0.1652). So there is no perfect orthogonality but there exists low correlation. This could be supported by the graph of the log quotients of the aforementioned groups of variables (Fig. 7).

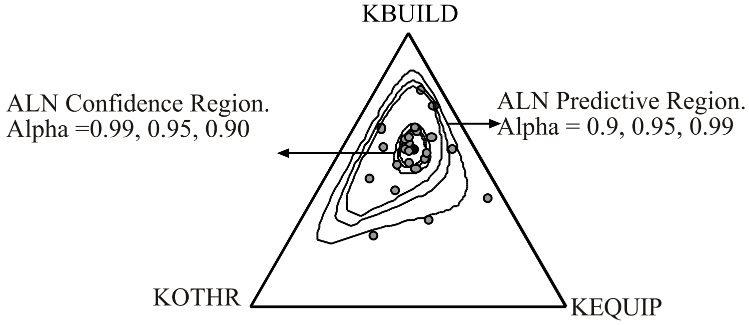
.

Figure 7. Low negative correlation between group of components



Now the variability has been reduced if we observe the amalgamated variables in the simplex in Figure 8 (total variance of 1.41988). We defined KBUILD = KNRES + KRES, and KEQUIP = KDUR + KTRAN. The names of the amalgamated variables have to do with KNRES and KRES representing building industries and KDUR and KTRAN representing factories or manufacturing sector of some kind of equipment.

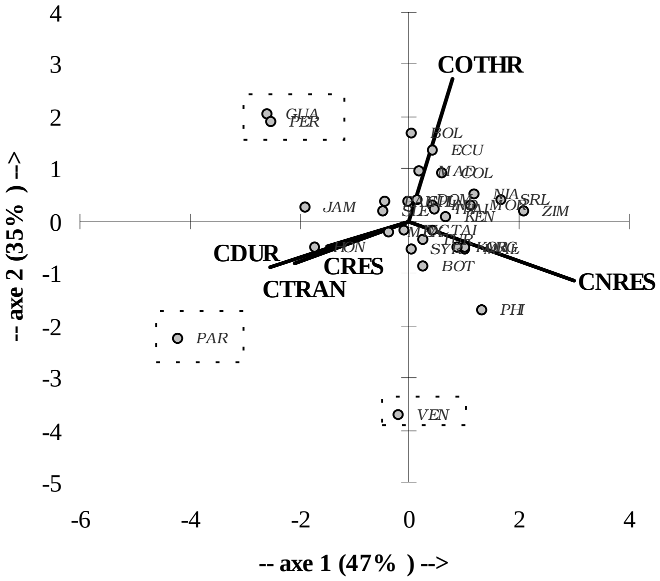
Figure 8. Ternary diagram for High income sample amalgamated variables and ALN confidence and predictive regions



As seen in Figure 8, variability has been slightly reduced. In any case, the additive logistic normal confidence (ALN) regions do not capture very well the data and predictive regions capture very well most of data except for 3 outliers.

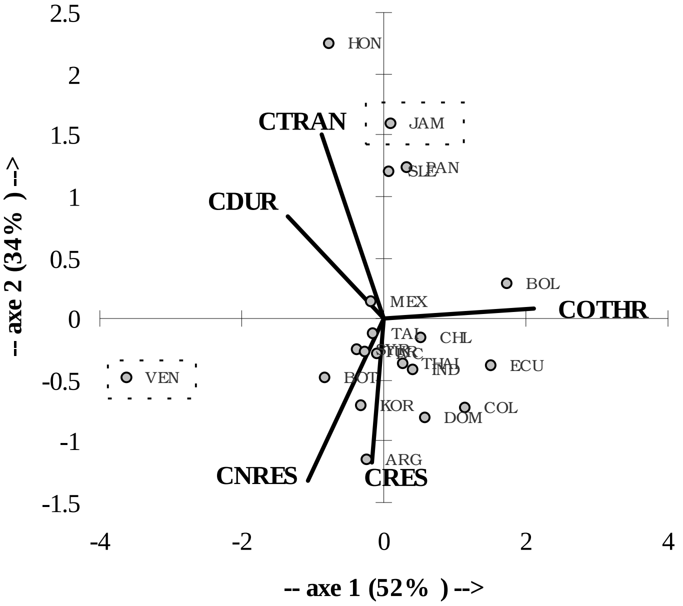
Now we proceed with the low income sample. By estimating its first two principal components we could recall the results of the full sample estimation in Figure 2 where the three components CDUR, CTRAN and CRES behaved coincidentally (Fig. 9). The low income sample includes most of countries included in the analysis after the zero replacement strategy was applied. Because it could be suspected that this data could act in the process as an outlier, we proceed to estimate again principal components but excluding the countries with zeroes in their data. Venezuela (VEN), Paraguay (PAR), Guatemala (GUA) and Peru (PER) also seem to behave as an outliers as indicated by the pointed boxes based on atypicality indices. The last three countries belong to the zero replaced countries (Table 4).

Figure 9. Biplot of first two principal components (82%) – Low income sample



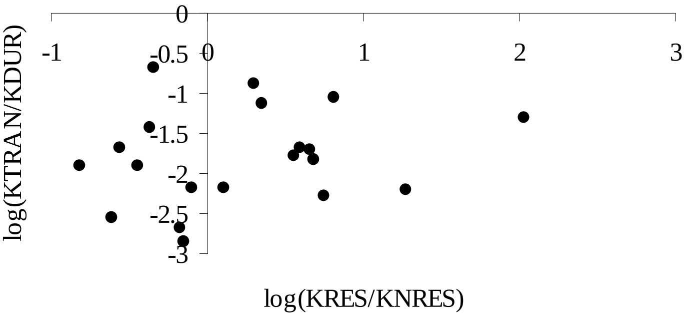
The results of this new estimation show that this filtered sample estimation behaves in a similar way that of the high income sample estimation. Again, as in Figure 6, CTRAN and CDUR show coincident vertices similar as CRES and CNRES (Fig. 10). It is remarkable the negative correlation between CTRAN and CRES (*r* = -0.5234772). This way we can make a step forward in the identification of the countries with zeroes in the data as outliers or at least members of a different population than the average under study. Still remains two potential outliers: Jamaica (JAM) and Venezuela (VEN) detected by atypicality indices.

Figure 10. Biplot of first two principal components (86%) – Low income sample without 'zero' countries



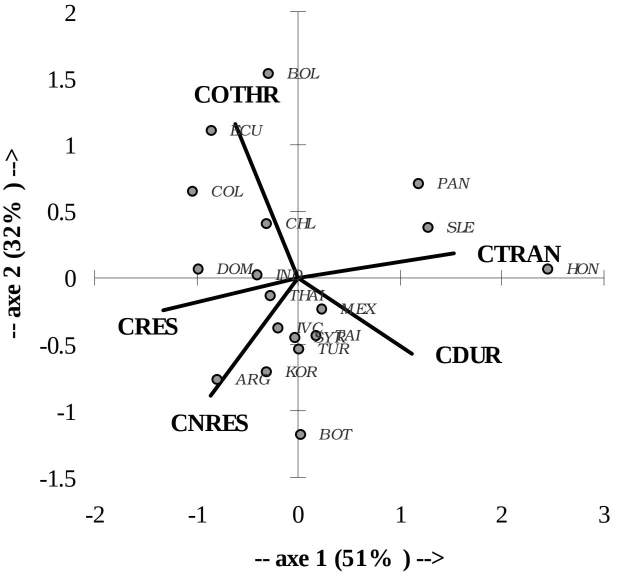
As in the former case we now proceed to show the low correlation present in the two log quotients observed in the Figure 10. In this case *r* = 0.2051 and we can see the scarce correlation present in this case by observing Figure 11.

Figure 11. Low correlation between group of components



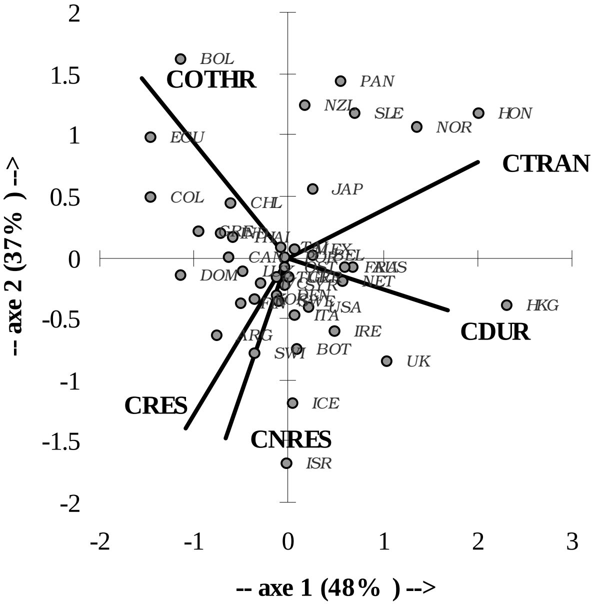
How about the presence of Venezuela and Jamaica as outliers? We can estimate again and watched that this relationship holds and is not very sensitive to presence of extreme cases (Fig. 12). Now it is more remarkable the negative correlation between CTRAN and CRES (*r* = -0.7194457) while CNRES and CDUR are markedly orthogonal.

Figure 12. Biplot of first two principal components (80%) –
Low income sample without 'zero' countries and Venezuela and Jamaica (outliers)



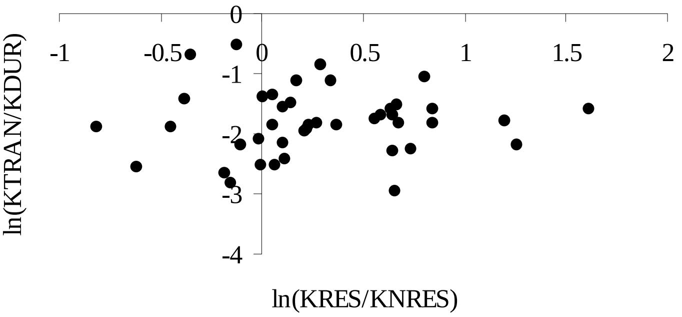
Because of this new finding we estimated again the full sample but discarding the countries with zeros in their components. This would confirm the outlier behavior of these countries observed previously in this paper. Figure 13 shows the biplot for this estimation. As suspected, relationship displayed in Figure 13 is extremely close to that observed in high income countries and low income countries without countries with zeros in their components. Again, construction related components behaved similarly and the something less clear happens with equipment related components. Durable goods component shows negative correlation ($r = -0.642377$) with other buildings capital proportion and transportation equipment component is also negative correlated ($r = -0.623674$) but with the amalgamated component of building sector (KBUILD). It seems that countries when assign capital to building sector at the same time they resign capital previously allocated in the transportation equipment sector. In the same line of reasoning, when countries allocate capital in producing durable goods, they sacrifice other kind of buildings investments.

Figure 13. Biplot of Full sample without ‘zero’ countries and outliers (85%)



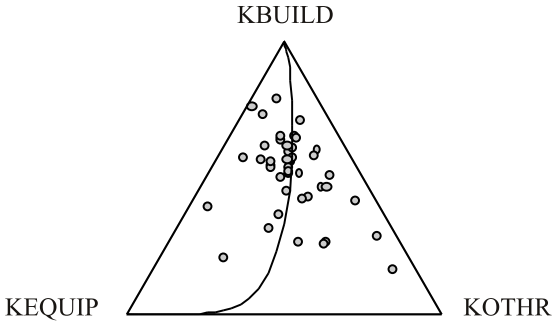
Similarly to Figure 11 and Figure 7, Figure 14 displays that the correlation between the pair of components is even smaller (*r* = 0.045277).

Figure 14. Low correlation between group of components



Now data got close to a compositional straight line that passed through the baricenter of the sample (Fig. 15) while data remain with a high variability (total variance of 7.87485).

Figure 15. Ternary diagram for Full sample amalgamated variables without 'zero' countries and outliers jointly with a compositional straight line



4.3 A note on the presence of zeroes and the quality of data

The exclusion of the zero data replaced countries shows that there was an underlying pattern on the variables that the presence of these countries hide from the analysis. Once we excluded them, the conclusion of the analysis seems to be very similar whether we treat with high income or low income countries.

What countries provoke this disturbance? The majority of the countries that have zeroes in several of the time series are low income countries and most of the them, 7 of 12, are African countries at the bottom of the income level category. Poor countries have deficient to non-existent statistical offices or measurement infrastructure. Data collected by Penn World Table authors and World Bank officers rely on country governments to provide local data for designing their indices[3]. Social and economic scientists, at the end of the provision line, shall trust in the quality of these data for making their research. So this is not the case of other empirical sciences where the researcher can, up to some point, control the quality of the data for her experiment or research work.

Due to this problem, the estimation proposed by this research should differentiate these poor countries from the rest of the sample and treat them, in some cases, as outliers or, in other cases, as a another population, subject to a particular study.

The others outliers present in the sample are particular cases that has been extracted from the sample in the last estimations for statistical purposes. The validity of this procedure is open to discussion.

5. Preliminary Conclusions and Discussion

We analyzed a static sample of capital per worker composition trying to understand the internal compositional changes that have taken place into the sample. We distinguished main patterns of behavior as follows:

Capital components from the income-based clustering showed a different behavior if we consider the countries with zeroes in their components. Without considering these data the biplot displayed a similar behavior either in the full sample or in subsamples estimations.

We identified two pairs of components that are highly correlated. Interestingly, they both refer to a same economic sector. Components related to residential and non residential buildings, in one hand, and

[3] Hofmann (1980) pointed out the seriousness of the problem for planning accurate development international programs. Without quality and quantity data for the precise estimation of endowments and shortage of resources in each undeveloped country, international programs could miss to help the countries in state of need.

components related to durable goods manufacturing and transportation equipment, in the other hand, behaved in remarkably correlated way.

Countries with zeroes in their components mostly belong to the lowest income category. This probably reflects another population different from the rest of the sample. In fact, after excluding 'zero' countries still remain potential subpopulations in the sample together with potential outliers, but estimation seems less sensitive to their presence. All these observations are supported by the continuos presence of high variability in the data.

We conjecture two possible explanations for the observed behavior. First, displacement among sectors (presence of collinearity) could be interpreted as a sectorial struggle for capital allocation. Assigning capital to one sector necessarily implies diminishing capital to another. This report helps to see the direction and affected sectors of these changes. Second, coincident vertices show sectors that exhibit a joint behavior between them: they raise and fall together during the economic process. The observed case of KBUILD (KRES + KNRES) could be better understood as the behavior of two complementary sectors: this way, following Canning and Bennathan (2001) observations on the externality approach to infrastructure research, increment in non-residential construction is made jointly with an increment in the residential counterpart (the dam and the required workers' houses initially exemplified). This is much less clear in the equipment sector.

Transportation and durable goods show less correlated between them and durable goods component behaved negatively with other kind of buildings. In the case of full sample and high and low income subsamples the relationship shows, following De Long and Summers (1990) findings, a sacrifice of other kind of building investment by increasing manufacturing participation in their stock of capital. At the same time, the building sector (residential and non residential) behaves negatively with transportation equipment when observing the full sample and high income subsample behavior. In the case of low income subsample, the displacement of transportation equipment participation is by reducing only residential building investment. Broadly speaking, whether we make durable goods, we resign other kind of buildings. Whether we build, we resign transportation equipment in the process.

We could mention as future paths of research two main approaches: First, we worked only with a very limited quantity of components and subjects. It would be desirable to analysis a higher number of countries and components to make more accurate conclusions. Second, there's no dynamical analysis in this report. It would be interesting to consider how these patterns have changed over the sample period. This could bring some evidence on potential structural breaks or sudden changes in the capital composition over time. Finally, and especially related with the former proposition, it could be highly motivating the study on how capital composition has influenced the economic growth process. For this purpose, it would be interesting to test this relationship using the currently available and extensive growth empiric datasets and research papers.

Acknowledgments

I would like to thanks Vera Pawlowsky-Glahn for comments and corrections to a very early draft and Carles Barceló-Vidal for extremely useful corrections made it to a previous version. The tests for logistic normality of the sample were conducted using MVN freeware by Andries van den Ark, Peter van der Heijden, and Ab Mooijaart. Compositional data analysis was conducted using Codapack software freely provided by Santiago Thió Fernández de Henestrosa. This paper was finished while I was invited researcher in the Department of Economics at Universidad Empresarial Siglo 21, Córdoba, Argentina, and I am particularly grateful to Marcelo Delajara. I am also grateful to Fernando Tohmé for his permanent support and advise and to the National Council for Scientific and Technological Research (CONICET) for their financial support. All errors are mine.

References

Aitchinson, J. (1986), The Statistical Analysis of Compositional Data, Chapman and Hall Ltd., London, 416 p.

Aitchison, J. (1997), "The one-hour course in compositional data analysis or compositional data analysis is easy", in Proceedings of the Third Annual Conference of the International Association for Mathematical Geology (ed. V. Pawlowsky-Glahn), 3-35, CIMNE, Barcelona.

Aitchinson, J. and M. Greenacre (2000), "Biplots of Compositional Data", *Journal of the Royal Statistics Society*, Series C, Part 4, *Applied Statistics* 51, 375-392.

Barceló-Vidal, C., J.A. Martín-Fernández, and V. Pawlowsky-Glahn (2001), "Mathematical Foundations of Compositional Data Analysis", 2001 Annual Conference of the International Association for Mathematical Geology Papers, Kansas Geological Survey, University of Kansas.

Billheimer, D., P. Guttorp , and W.F. Fagan (1998), "Statistical Analysis and Interpretation of Discrete Compositional Data", *NCRSE Technical Report Series* No. **011**.

Brandt, P.T., B.L. Monroe, and J. T. Williams (1999), "Time Series Models for Compositional Data", Proceedings of the Workshop in Political Theory and Policy Analysis, July 7, Indiana University, Bloomington, USA.

Brehm, J., S. Gates, and B. Gomez (1998), "Donut Shops, Speed Traps, and Paperwork: Supervision and the Allocation of Time to Bureaucratic Tasks", Proceedings of the Annual Meeting of the Midwest Political Science Association, April 23-25, Chicago, USA.

Canning, D. (2000), "The Contribution of Infrastructure to Aggregate Output", Working Paper Series **2246**, World Bank, Washington.

Canning, D. and E. Bennathan (2000), "The Social Rate of Return of Infrastructure Investments", *Policy Research Working Papers* **2390**, World Bank, Washington.

De Long, B. and L. Summers (1990), "Equipment Investment and Economic Growth", Working Paper **3515**, National Bureau of Economic Research.

Devarajan, S., V. Swaroop, and H. Zou (1996), "The composition of public expenditure and economic growth", *Journal of Monetary Economics* 37, pp. 313-344.

Fry, J.M., T.R.L. Fry, and K.R. McLaren (1996), "Compositional Data Analysis and Zeros in Micro Data", *Applied Economics*, Vol. **32**, 2000, pp. 953-959.

Glaeser, E.L. and J. Gyourko (2001), "Urban Decline and Durable Housing", Working Paper **8598**, National Bureau of Economic Research, MA.

Guasch, J.L. and J. Kogan (2001), "Inventories in Developing Countries: Levels and Determinants, a Red Flag on Competitiveness and Growth", Working Papers Series **2552**, World Bank, Washington.

Henderson, J.V., Z. Shalisi, and A.J. Venables (2002), "Geography and Development", World Bank Group, mimeo.

Hofmann, H. (1980), "Statistics in the Third World: Problems and Perspectives", *Economics*, Institute for Scientific Co-Operation, Tübingen, Germany, Vol. **21** pp. 100-115.

Ingram, G. and Z. Liu (1997), "Motorization and Road Provision in Countries and Cities", Working Paper Series **1842**, World Bank, Washington.

Ingram, G. and Z. Liu (1999), "Determinants of Motorization and Road Provision", *Policy Research Working Paper* **2042**, World Bank, Washington.

Jones, C. (1994), "Economic growth and the relative price of capital", *Journal of Monetary Economics* **34**, No. 3, December, pp. 359-382.

Jovanic, B. and R. Rob (1997), "Solow vs. Solow: Machine Prices and Development", Working Paper **5871**, National Bureau of Economic Research, Cambridge, MA.

Klingebiel, D. and J. Ruster (2000), "Why Infrastructure Facilities Often Fall Short of Their Objectives", Working Paper **2358**, World Bank Institute, Washington.

Martín-Fernández, J.A., C. Barceló-Vidal and V. Pawlowsky-Glahn (1998), "A Critical Approach to Non-Parametric Classification of Compositional Data", Proceedings of the 5th Conference of the International Federation of Classification Societies, Universitá La Sapienza (Rome, Italy)..

Martín-Fernández, J.A., C. Barceló-Vidal and V. Pawlowsky-Glahn (2000), "Zero Replacement in Compositional Data Sets", Proceedings of the 7th Conference of the International Federation of Classification Societies (Namur, Belgium).

Pearson, K. (1897), "Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs", Proceedings of the Royal Society **60**, 489-498.

Randolph, S., Z. Bogetic, and D. Heffley (1996), "Determinants of Public Expenditure on Infrastructure: Transportation and Communication", Working Paper Series **1661**, World Bank, Washington.

Reinikka, R. and J. Svensson (1999), "How Inadequate Provision of Public Infrastructure and Services Affects Private Investment", *Policy Research Working Paper* **2262**, The World Bank, Washington.

Seitz, H. (1995), "Public capital and the demand for private inputs", *Journal of Public Economics* 54 (2), 161-324.

Seitz, H. (2000), "Infrastructure, Industrial Development, and Employment in Cities: Theoretical Aspects and Empirical Evidence", *International Regional Science Review* 23, 3 (July): 259–280.

Temple, J, and H.S. Voth (1998), "Human capital, equipment investment, and industrialization", *European Economic Review*, July, 42(7), 1343-1362.

Turnovsky, S.J. and W.H. Fisher (1995), "The composition of government expenditure and its consequences for macroeconomic performance", *Journal of Economic Dynamics and Control* 19, pp. 747-786.

World Bank (2000), "World Bank Development Indicators 2000", The International Bank for Reconstruction and Development, The World Bank, Washington, DC.

APPENDIX

Table 4. Full sample raw data and income categories

Country	KTRAN	KOTHR	KDUR	KRES	KNRES	Zero presence	Income category
ARG	0.012461354	0.184198133	0.083301619	0.28025142	0.439787474		3
AUS	0.055392462	0.173185373	0.219130611	0.283768304	0.26852325		4
BEL	0.035461809	0.226364254	0.223986369	0.292061381	0.222126187		4
BOL	0.011530349	0.75326785	0.062197221	0.111274655	0.061729925		2
BOT	0.01612742	0.12913198	0.275897355	0.267852481	0.310990764		3
CAN	0.020162395	0.29137834	0.094665541	0.392439285	0.201354438		4
COL	0.00978964	0.510051236	0.057581989	0.268494913	0.154082222		2
CHL	0.02636323	0.380615886	0.076238766	0.357168588	0.159613531		3
DEN	0.023967068	0.202923089	0.165103874	0.338195439	0.26981053		4
DOM	0.009677825	0.293623183	0.08772119	0.474727144	0.134250658		2
ECU	0.009652085	0.641886605	0.05243228	0.19443052	0.10159851		2
FIN	0.013957561	0.227184123	0.173058375	0.302267958	0.283531983		4
FRA	0.050113752	0.178993963	0.223657494	0.292748015	0.254486775		4
GER	0.027849073	0.217104832	0.181540981	0.319718118	0.253786997		4
GRE	0.012247335	0.393372306	0.122148019	0.309615522	0.162616819		4
GUA	0.012441894	0.486649946	0.268815511	0.227074307	0.005018342	1	2
HKG	0.134604767	0.052228043	0.413830561	0.216663986	0.182672643		4
HON	0.174410167	0.19448265	0.417390485	0.122351308	0.091365391		1
ICE	0.018383095	0.07113572	0.112446204	0.612932131	0.185102851		4
IND	0.015476006	0.372038977	0.135968004	0.251029738	0.225487274		1
IRE	0.033647684	0.121533233	0.220621884	0.320583286	0.303613913		4
ISR	0.009800368	0.053936131	0.19159013	0.491306086	0.253367286		5
ITA	0.026442129	0.152115236	0.166886705	0.458214583	0.196341347		4
IVC	0.018726354	0.230669373	0.182807511	0.384289754	0.183507009		1
JAM	0.07162015	0.287088626	0.263995919	0.333374186	0.043921119		2
JAP	0.046401373	0.33477738	0.190025594	0.214825131	0.213970522		4
KEN	0.006024031	0.247744536	0.167350986	0.353397025	0.225483422	1	1
KOR	0.017793609	0.205790335	0.120179689	0.201109994	0.455126373		3
LUX	0.015635503	0.27766447	0.179020257	0.279620572	0.248059198		4
MAD	0.007075487	0.471981513	0.262368523	0.151847709	0.106726768	1	1
MAL	0.007011287	0.164628763	0.230489623	0.195505738	0.40236459	1	1
MEX	0.031539465	0.247502162	0.196447831	0.348348087	0.176162455		3
MOR	0.005013717	0.290097294	0.08200957	0.351163252	0.271716167	1	2
NET	0.045949068	0.165861367	0.220472568	0.298300143	0.269416854		4
NIA	0.005964026	0.397770275	0.103536076	0.209001225	0.283728398	1	1
NOR	0.145018143	0.283609504	0.250553733	0.15073494	0.170083681		4
NZL	0.0415393	0.486192866	0.204156697	0.187531832	0.080579304		5
OST	0.024772457	0.241835217	0.203867119	0.262760624	0.266764582		4
PAN	0.078678776	0.490851569	0.156380896	0.113197803	0.160890957		3

PAR	0.05023068	0.005018377	0.144757772	0.794974794	0.005018377	1	2
PER	0.011397385	0.387038855	0.111814088	0.484728192	0.00502148	1	2
PHI	0.005025157	0.043552756	0.147520273	0.219935539	0.583966275	1	2
POR	0.028878586	0.208860152	0.141421831	0.517494764	0.103344667		4
SLE	0.06284731	0.40218092	0.2625121	0.11090145	0.16155822		1
SPA	0.006510182	0.265627998	0.069003067	0.55679444	0.102064313	1	4
SRL	0.008341784	0.389148471	0.048702067	0.124560715	0.429246964	1	1
SWE	0.024671546	0.191327703	0.158986524	0.37037939	0.254634837		5
SWI	0.013369418	0.152773213	0.169630939	0.331904968	0.332321462		5
SYR	0.035039572	0.197077915	0.107324286	0.38630771	0.274250517		2
TAI	0.019326469	0.272855523	0.248799516	0.161091288	0.297927204		3
THAI	0.01349893	0.352324158	0.19618864	0.199517106	0.238471165		2
TUR	0.022054802	0.232156999	0.196500907	0.261155642	0.28813165		3
UK	0.041568107	0.073740364	0.298971922	0.32416378	0.261555827		4
USA	0.032885528	0.157020487	0.164615814	0.421873622	0.22360455		4
VEN	0.035446795	0.006914265	0.187064831	0.278381186	0.492192924		3
ZIM	0.005025618	0.288908179	0.042031064	0.114431076	0.549604063	1	1

Income categories: 1. low income, 2. lower middle income, 3. upper middle income, 4. high income (OECD countries), 5. high income (non-OECD countries)

Table 5. Full sample descriptive statistics

Descriptors	KTRAN	KOTHR	KDUR	KRES	KNRES
Min	0.005013717	0.005018377	0.042031064	0.11090145	0.005018342
Max	0.174410167	0.75326785	0.417390485	0.794974794	0.583966275
Mean	*0.031157859*	*0.263535583*	*0.173852132*	*0.301406658*	*0.230047768*
Standard dev.	0.033864678	0.151133169	0.080697053	0.135227302	0.126450691
Median	0.019744432	0.236996108	0.171344657	0.287914842	0.225485348
Std dev. of mean	0.004525358	0.020196019	0.010783597	0.018070508	0.016897684
Sdm/Mean	0.145239698	0.076634885	0.062027409	0.059953912	0.073452938
Kurtosis	7.912725563	1.199400182	1.230337554	2.225203047	0.7857099
Skewness	2.676145409	0.832780102	0.742028075	1.119888708	0.643724523

N = 56.

Table 6. Full sample and subsamples component averages

Income category	KTRAN	KOTHR	KDUR	KRES	KNRES
low income	0.009205574	0.320361261	0.146656732	0.223007872	0.300768561
lower middle income	0.029194012	0.330761911	0.153553023	0.327897408	0.158593646
upper middle income	0.028865769	0.238890761	0.171201268	0.252061832	0.30898037
high income OECD	0.044259129	0.20998688	0.192719103	0.342443935	0.210590953
high income nonOECD	0.021351775	0.225541542	0.183548297	0.324358476	0.245199909
full sample	0.030581687	0.261014759	0.172240133	0.304870389	0.231293032