Online Data Appendix for:

"Buyer-Seller Relationships in International Trade: Evidence from U.S. State Exports and Business-Class Travel"

Anca Cristea
University of Oregon

December, 2010

The Online Data Appendix is organized in two parts.

Part I provides a detailed description of the steps involved in constructing the restricted sample of international travel and trade flows used in the empirical part of the paper. The focus is primarily on the treatments applied to the original air travel dataset DataBank 1B (DB1B) Passenger Origin and Destination Survey, provided by the US Department of Transportation (DOT). While this dataset is new to the trade literature, it has been previously employed in empirical industrial organization studies (Brueckner, 2003; Whalen, 2007, and references therein), so my approach in preparing the air travel data follows in many respects previous practice.

Part II compares the air traffic measures computed from the DB1B dataset with corresponding representative values from the T100 Market dataset in order to gauge the likelihood of measurement bias in the estimation sample.

PART I: Construction of Estimation Sample

1 Overview of the Air Travel Dataset

The DB1B is based on a 10% quarterly sample of domestic and international airline tickets, where at least one flight segment is services by a US air carrier or a foreign carrier that is part of an aviation alliance under US antitrust immunity. The dataset is unique in that it provides air fare as well as true origin and destination information at a very detailed level. A record in the DB1B corresponds to a an airline ticket and indicates the complete flight itinerary at airport detail (including the return route for round trip tickets), the number of passengers included on the ticket, distance traveled, airfare paid, flight class type (e.g., first, business, economy), the trip break points along the itinerary[1] (i.e., airports where the passengers interrupt the trip rather than take a connecting flight), and other ticket characteristics.

[1]For example, in the DB1B data an entry for a round trip ticket can indicate the following itinerary: Chicago OHare - New York City JFK - London Heathrow - New York City JFK - Chicago OHare. To identify the airports where the travelers interrupt their journey, the dataset includes a trip break indicator variable. In this example, London Heathrow could be the trip break point. But in more circuitous tickets, more than one airport can be signaled as a trip break point.

2 Preparing the Air Travel Sample

The raw dataset is restricted in several ways to conform to the papers empirical objectives and also to reduce the incidence of coding errors.

1. I drop all the airline tickets that correspond to domestic flights or international flights that only transit the US. This way I can focus only on international flights that either depart (outbound) or arrive (inbound) in one of the contiguous US states.

2. I drop the circuitous itineraries, defined as tickets that have more than one trip break point per itinerary. This is done to avoid difficulties in assigning one itinerary to a unique bilateral origin-destination pair. The single trip break point for each remaining ticket is then used to identify the destination of the travelers.

3. To reduce the incidence of coding errors in ticket prices, I remove the fare information from several observations.[2] First, I remove the fare information for all the tickets whose prices are marked as unreliable by the assigned DOT indicator. Following criteria commonly used in previous studies, I also remove fare information for tickets with price values below $100 as well as those with prices outside the range 1/4 to 4 times the geometric average fare for a US state-foreign country route. Finally, I remove the fare and distance information from highly unusual tickets that have more than eight flight segments per itinerary (respectively more than four flight segments for one-way itineraries). After cleaning the air fare variable of noisy values, I define the ticket price as a single-direction fare. So, the price for one-way flights remains exactly as listed in the original data, whereas the fare for round-trip flights is replaced with one-half the value listed in the DB1B data. This is done in order to have fares that are comparable across airline tickets. I apply the same procedure to the ticket distance variable, in order to get single-direction distances across all tickets.

After filtering the ticket data for each quarter of the DB1B, I combine the tickets from all 24 quarters into one single panel sample for the period 1998-2003. The resulting dataset includes between 2.2 to 2.4 million airline ticket observations per sample year. I then use a DOT concordance (amended with US Census country codes) to assign to each ticket's origin and final destination airport codes the corresponding US state and foreign country respectively.

In addition, I assign each US state to a larger US region. Clustering neighboring states into a region is necessary because many large international airports are sufficiently close to a states borders to be able to serve out-of-state air travelers. The allocation of states to regions results in 17 US aviation regions, which are created based on the following two criteria: states that share access to a large gateway airport are grouped together, and each region must include at least one large hub or major gateway airport (the allocation of states to regions is presented in the paper Appendix Table A1) [3] Some foreign countries in the sample are also grouped into larger world geographic regions (generally small and less developed countries). The need to cluster foreign countries into world regions is dictated by the format of the original foreign-born population dataset provided by the U.S. Census.[4]

[2] I do not drop from the sample the record entirely because it can still bring information about other ticket characteristics that are less noisy (such as the number of travelers). However, dropping these observations entirely would not change the results.

[3] The classification of airports is provided by the Federal Aviation Administration (FAA).

[4] In the US Census state level data on foreign born population by country of origin, an observation could indicate the number of German born people living in Ohio, but also the number of people born in 'Other Western Africa' countries living in Ohio. A concordance is provided by the US Census to precisely indicate which are the countries included in the geographical region 'Other Western Africa'.

3 Data Aggregation and Variable Construction in the Restricted Air Travel Sample

Using the resulting airline ticket dataset, I create several new variables that are needed for my econometric purposes. First, I construct an indicator variable for the direction of air travel, to distinguish between outbound flight tickets (i.e., itineraries that originate in the US and have the final destination abroad) and inbound flight tickets (i.e., itineraries that start in a foreign country and arrive at a destination in the US).Then, I create an indicator variable for round trip tickets, defined as itineraries that originate and terminate in the same city. Finally, since in the original DB1B dataset the flight class type is specific to each flight segment of an itinerary, I create an indicator variable that assigns the class type – business or economy – to the entire travel itinerary. I consider as business class any airline ticket that has a distance-weighted fraction of business/first class flight segments greater than one half. That is, I compute the following statistic:

$$business\_class = \sum_{s=1}^{S} \left(\frac{\text{segment dist}_s}{\text{total ticket dist}} \right) \cdot I_s(1 = \text{business/first class}) \tag{1}$$

where s indexes a flight segment and S is the total number of flight segments of a given airline ticket. If business_class ≥ 0.5 (i.e., more than 50% of the trip distance is flown at business or first class), then the itinerary is considered a business class ticket.[5]

After creating these additional variables, I can now dispense of the ticket level and quarter details of the original DB1B dataset by collapsing the data into US region-destination country-year observations, separately by direction of air travel (inbound and outbound) and by ticket class type (business and economy). Flight distances and air fares are computed as passenger-weighted averages. Air fares are also deflated by the US GDP deflator in order to be expressed in constant US dollars. I create separate samples for outbound and inbound air travel flows. An observation in the resulting outbound sample corresponds for example, to business class air travel in year 2000 departing from the US Great Lakes region to arrive to Japan and indicates the total number of business class travelers[6], the average business class air fare and the average business class trip distance, combined over the one-way and round-trip flights (as long as they have the same origin region and foreign destination country).

4 Merging the Air Travel Sample with the US Exports Dataset

The final step is to combine the restricted air travel dataset with the other data sources used for the empirics, among which the US manufacturing exports data is the most important.

The state level exports data is available at 3-digit NAICS classification codes from the Origin of Movement series provided by the US Census. Before merging it to the aviation data, the raw trade dataset was adjusted as follows. I first remove from the original dataset agricultural sectors and focus only on manufacturing trade (i.e., NAICS codes 311 to 339). This is done in order to increase the likelihood that the states which are observed as shipping goods abroad are the same as the states of production. I then attach to each US state the corresponding aviation region (described previously), and also attach to each selected destination country the corresponding world region (following the same geographical regions defined in the foreign born population data). I collapse

[5]This definition of business class tickets is more restrictive than computing the simple fraction of segments traveled at business class, which is what has been used in the industrial organization literature (Brueckner (2003) among others).

[6]The number of travelers is going to be measured in multiples of 10, as the original data is a 10% sample.

the bilateral export flows by sector and by US origin region-destination country-year pair. Then I compute for each manufacturing sector the fraction of differentiated 4 digit SITC codes in a given 3-digit NAICS category, using the Rauch (1999) 'liberal' classification of goods and the Feenstra-Lipsey concordance between SITC and NAICS categories, as provided by the NBER. For each US origin region-destination country-year record I compute the degree of differentiation in bilateral exports as follows:

$$Export\_Composit_{sj} = \sum_{h} \theta_h \frac{X_{sjh}}{X_{sj}} \tag{2}$$

where h indexes a 3-digit NAICS manufacturing sector with positive exports shipped from US region s to destination country j; θ_h is the information intensity of a sector and is measured as the fraction of Rauch differentiated goods in a given sector; and X denotes the volume of manufacturing exports. Finally, I can now dispense of the sector detail and collapse export values across sectors by US region-destination country-year. So, in the resulting sample an observation corresponds to an origin region-destination country-year pair and indicates the total value and the degree of differentiation of manufacturing exports.

Now the two main datasets on bilateral outbound (inbound) air travel and export flows can be merged together. The merge is realized by US region-destination country-year and the results are described in the paper Appendix Table A2. While the merge is not exact (mostly because of missing business class air travel flows), the bilateral pairs that are dropped correspond to a very small share of no more than 0.5% of total US manufacturing exports by value. This ensures that the resulting sample of bilateral trade and travel flows is representative of US exports. Adding the remaining control variables to the restricted sample is straightforward: the variables require no initial preparation (other than collapsing the original data to the corresponding US region- foreign country-year level of aggregation) and it is generally a precise merge. The sources for the control variables are: the 2000 US Decennial Census for state level data on foreign born population by country of origin; the Bureau of Labor Statistics (BEA) for state GDP and employment level in foreign affiliates by country of ultimate beneficiary owner; the 2002 US Economic Census for the sector level Herfindahl-Hirschman Index available at 3-digit NAICS categories.

PART II: Representativity of the DB1B data: Evidence from Comparison with the T100 Market data

The DB1B Passenger Origin-Destination Survey is not a representative sample of international airline tickets, since it only reports data collected from U.S. air carriers or immunized foreign carriers. The dataset omits passengers that travel on flights operated exclusively by un-immunized foreign carriers, and so the constructed international air travel flows are under-represented. To get a sense of how severe this sampling error is, I compare the DB1B dataset with a representative air traffic database – T100 International Market – provided by the U.S. Department of Transportation.

The *T100 International Market* is a firm level dataset that provides information on capacity and air traffic volumes for all cross-border online flight itineraries operated by both domestic and foreign air carriers. The data is collected monthly at carrier–origin airport–destination airport level, and reports for each carrier-route pair the number of onboard passengers per direction of travel. One important benefit of the T100 Market dataset is that it provides an exhaustive account of air passenger traffic crossing the U.S. border by route and operating carrier.

Table 1: Airline Market Shares in the T100 Data

	Share of International Air Passenger Traffic Supplied by:		
	U.S. Carriers (%)	*Foreign Carriers:*	
		with Immunity (%)	*No Immunity (%)*
1998	48.53	11.23	40.24
1999	47.85	12.36	39.79
2000	47.96	12.27	39.77
2001	48.67	13.27	38.06
2002	49.75	16.22	34.03
2003	50.34	16.62	33.04
Entire Sample	***48.85***	***13.66***	***37.49***

Table 1 summarizes the aggregate market share of U.S. and foreign air carriers in total cross-border air passenger transport, with the foreign airlines distinguished based on participation in antitrust immunity alliances. Two things are worth pointing out. First, the market share of US carriers – averaging 49% – has remained relatively constant over the sample period. Second, the share of international traffic operated by non-immunized foreign carriers is non-trivial, but it has dropped over time in favor of foreign carriers that are part of antitrust immunity alliances. Out of the passengers traveling on non-immunized foreign carriers only a fraction of them are not sampled in the DB1B ticket level dataset (i.e., those having *all* flight segments operated by non-immunized foreign air carriers). While it is not possible to evaluate more precisely the degree of miss-measurement in the DB1B dataset, at the end of the day what matters most for the econometric analysis is the extent to which this measurement error is random across aviation routes and orthogonal to the variables of interest, such as the patterns of U.S. state level exports.

Next, I proceed by merging the DB1B and T100 Market datasets in order to be able to compare the air passenger traffic flows from each source and gauge the severity of sampling bias in the DB1B dataset. Then, I examine whether the sampling bias is reduced, or maybe even completely eliminated, when conditioning bilateral air traffic flows on state and country-year fixed effects; that is, the same set of fixed effects employed in all the estimations reported in the main text of the paper. This analysis is done separately for U.S. outbound and inbound travel flows.

Merging the DB1B and T100 Market Datasets.The T100 Market dataset differs from the DB1B data in one important aspect: it does not report air traffic statistics based on passengers' true origin and destination. That is, the T100 Market dataset reports for a given origin-destination aviation market the number of passengers transported by each air carrier in the market during a time period, but with no information whether the passengers connect flights or begin their itineraries at origin, respectively whether they end their itinerary at destination or make further connections.

Given that the DB1B database reports the full itinerary of a sampled airline ticket, including all the connecting airports, I proceed by first breaking each sampled airline ticket into its component flight segments; then I aggregate air passengers by flight segment to obtain the volume of air traffic at city-pair market level (economy and business class traffic combined). Afterwards, I aggregate further the number of passengers across all airport pairs within a U.S. state by destination country, separately for each DB1B and T100 Market datasets, and then merge the two datasets. The resulting merged sample contains the DB1B air traffic measure at US state by foreign country

level, and the corresponding representative air traffic value from the T100 Market dataset.

Comparing the Two Datasets. Figure 1 illustrates the degree of measurement error in the DB1B outbound air traffic variable, as revealed from the direct comparison with the representative air traffic variable obtained from the T100 Market dataset.

If the volume of bilateral outbound air travel from the DB1B dataset were representative for the population of international airline tickets, all the data points should have been aligned on the 45 degree line. However, many data points lie dissproportionaltely below the 45 degree line, confirming the intuition that the outbound air travel flows in the DB1B dataset are under-represented due to the unreported airline tickets issued by un-immunized foreign air carriers.

To complement the information learned from the graph about the magnitude of the miss-measurement of air travel flows in the DB1B dataset, Table 2 reports correlation coefficients between the bilateral air travel flows computed from the DB1B and T100 Market datasets. The first thing that emerges from the table is that the correlation coefficients for total and for outbound air traffic are high, averaging 0.86 in each case. This implies that the great majority of bilateral travel data points are rather close to their true value, and that the under-represented flows observed in Figure 1 should not affect too much the main empirical results of the paper. The correlation coefficient is also high when restricting attention to long haul flights of more than 2000 miles per directional trip, but then it drops for thin markets with less than 3,000 outbound passengers per year, and also for thick markets with more than 25,000 passengers.

However, as previously stated, what is actually most relevant for the estimated results reported in the paper is whether the DB1B residual air traffic obtained after netting out the origin and destination fixed effects, still deviates in a systematic way from the population values. For this, I compute the residual air traffic level obtained after estimating the following regression model separately for each source of air travel data:

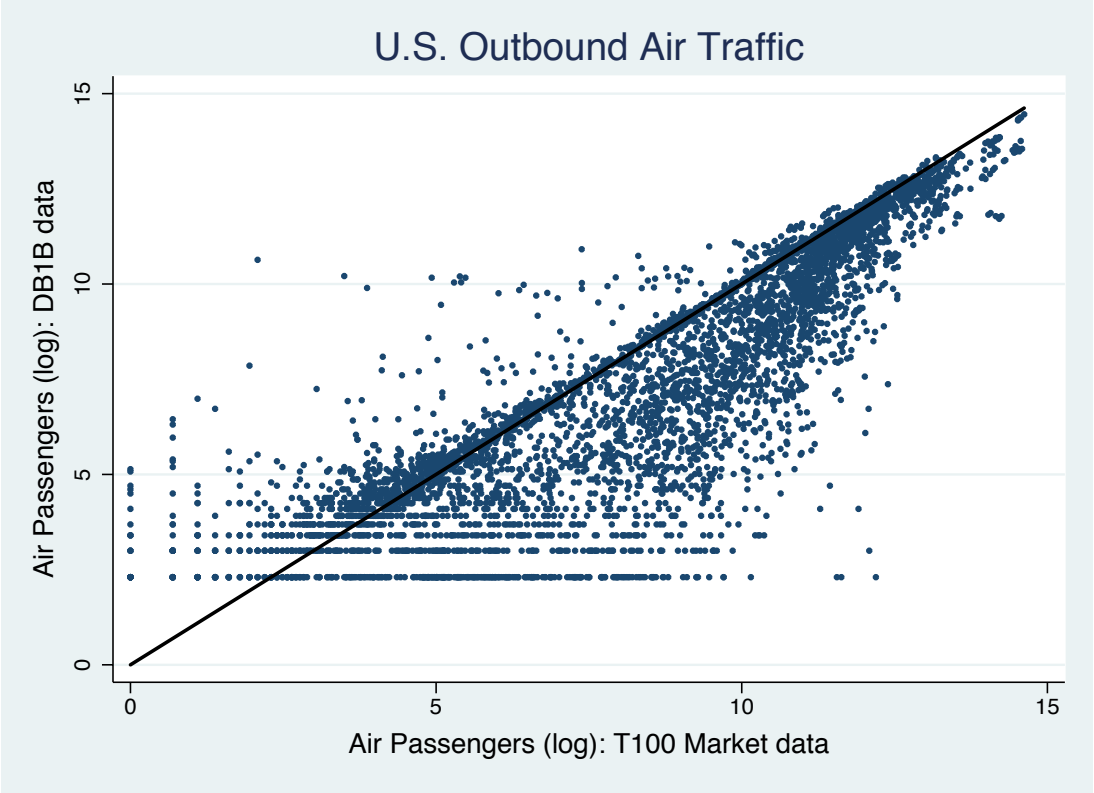
$$ln\text{Traffic}_{sjt} = \lambda_s + \lambda_{jt} + \epsilon_{sjt} \tag{3}$$

where s, j and t index the U.S. state, foreign country and year, respectively.

Figure 2 illustrates graphically the DB1B residual air travel flows and the corresponding residual values from the T100 Market dataset. The scatterplot reveals that by accounting for state and country-year fixed effects, the systematic measurement error component from the DB1B air traffic variable is purged. There still is measurement error left in the bilateral travel flows (as the data points in Figure 2 do not align perfectly on the 45 degree line), but if anything they should have a much smaller effect on the estimation results.

To further illustrate the correction in the sample bias, Table 3 reports correlation coefficients between the residual bilateral air travel flows as computed from the DB1B and T100 Market datasets. The coefficients are computed separately for outbound and inbound traffic. For comparison purposes, for each direction of travel I first report the correlation coefficients between the *actual* air traffic levels (i.e., dependent variable in equation (3)), and afterwards I report the correlation coefficients between the *residual* values. The results in Table 3 bring further confirmation that a large part of the error embedded in the constructed DB1B air travel flows is removed once accounting for the fixed effects structure of the baseline regression model in the paper. The fixed effects have the rectifying influence on very thin or very thick air traffic flows.

Figure 1: Comparison between the Number of Travelers: DB1B vs. T100 datasets

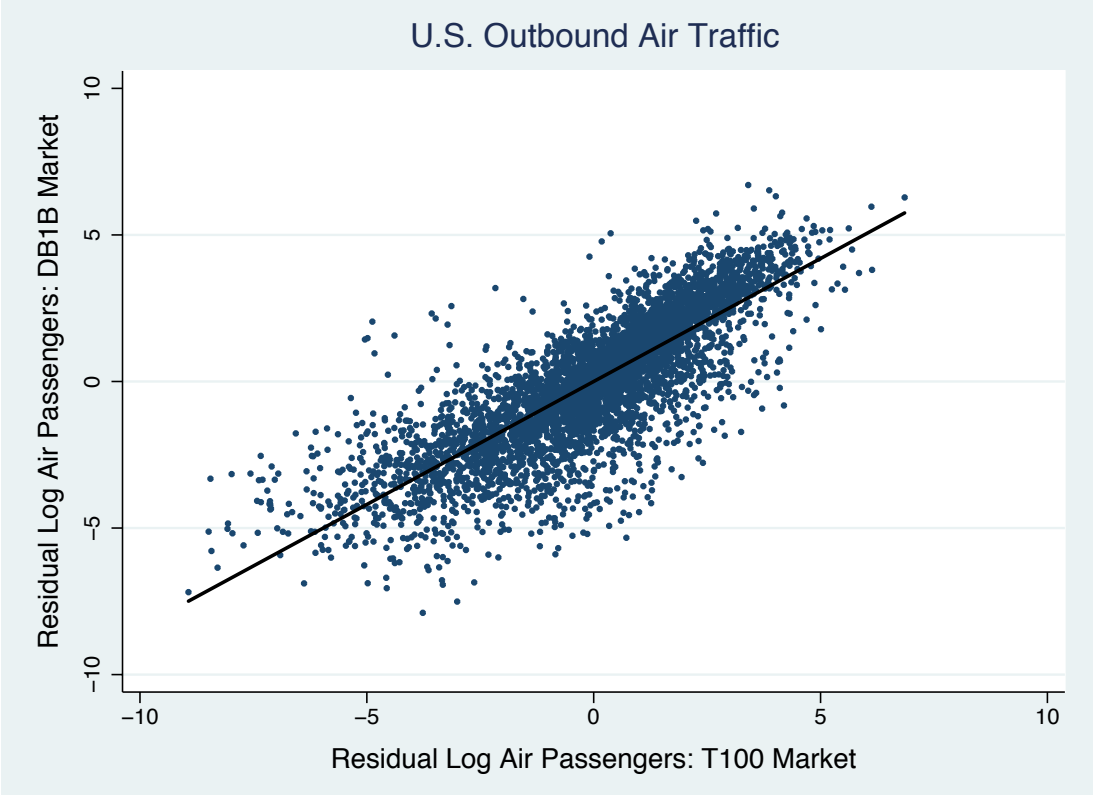


Note: The fitted line represents the 45 degree line.

Table 2: Correlation Coefficients between the Number of Travelers in DB1B vs. T100 datasets

	All traffic	Outbound Traffic			
year		all	long haul	thin mkt	thick mkt
1998	0.837	0.843	0.868	0.435	0.656
1999	0.844	0.847	0.871	0.401	0.635
2000	0.855	0.854	0.867	0.433	0.668
2001	0.867	0.868	0.874	0.418	0.676
2002	0.890	0.891	0.896	0.545	0.667
2003	0.874	0.874	0.868	0.526	0.655
Pooled	**0.861**	**0.863**	**0.874**	**0.458**	**0.659**

Figure 2: Comparison between the Residual Number of Travelers: DB1B vs. T100 datasets



Note: The fitted line represents the 45 degree line.

Table 3: Correlation Coefficients between the Number of Travelers in DB1B vs. T100 datasets

	Traffic			
	all	**long haul**	**thin mkt**	**thick mkt**
U.S. Outbound:				
traffic level:	0.863	0.874	0.458	0.659
traffic residual:	0.812	0.838	0.728	0.788
U.S. Inbound				
traffic level:	0.859	0.872	0.454	0.655
traffic residual:	0.815	0.843	0.732	0.791