Exaggerated Death of Distance

Revisiting Distance Effects on Regional Price Dispersions

Kazuko Kano[†]
School of Finance and Economics
The University of Technology Sydney
Haymarket, NSW
2001 Australia
Tel: +61-2-9514-3206
Fax: +61-2-9514-7711
Email: Kazuko.Kano@uts.edu.au

Takashi Kano
Graduate School of Economics
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku
Tokyo, 113-0033, Japan
Tel: +81-3-5841-5655
Fax: +81-3-5841-5655
Email: tkano@e.u-tokyo.ac.jp

Kazutaka Takechi
Faculty of Economics
Hosei University
4342 Aihara-machi, Machida-shi
Tokyo, 194-0298, Japan
Tel: +81-42-783-2566
Fax: +81-42-783-2611
Email: ktakechi@hosei.ac.jp

Current Draft: July 22, 2010
Very Preliminary
Comments Welcome

Abstract

Many past studies in the literature of the law of one price (LOP) show statistically significant but economically subtle roles of geographical distance on regional retail price differentials. In this paper, we challenge this empirical "death of distance" as a dominant source of violations of the LOP by making three contributions. First, this paper investigates a unique daily data set of wholesale prices of agricultural products in Japan that enables us to identify source regions of agricultural products and observe the daily delivery patterns of these products to consuming regions. Second, we build a simple structural model to explain the observed product-delivery patterns and claim theoretically that ignoring the underlying choice of delivery might result in a serious under-bias toward our inference on the role of distance in regional price dispersions due to sample selection. Finally, as the third contribution, we estimate a sample-selection model, which is imposed theoretical restrictions of our structural model, using the data of wholesale prices of several vegetables. Across all the vegetables this paper scrutinizes, we find large estimates of the elasticity of transportation costs with respect to geographical distance relative to the existing estimates. This paper, hence, provides evidence that conventional estimates of the distance elasticity could be heavily biased downwards and spuriously underestimate the role transportation costs play in regional price dispersions and LOP violations.

Key Words : *Law of one price; Regional price dispersion; Transportation cost; Geographical distance; Agricultural wholesale price; Sample-selection bias*

JEL Classification Number : F11, F14, F41

[†] We would like to thank Toni Braun, Daiji Kawaguchi, Michael Keane, Junko Koeda, Toshi Mukoyama, Makoto Nirei, Hiroshi Ohashi, Yasuyuki Sawada, Yi Chan Tsai, and seminar participants in the 2010 Asia Pacific Trade Seminar, Tokyo macro workshop (July 2010) and Hosei University for helpful and valuable discussions and suggestions. The first and second authors wish to thank the Kikawada foundation for financial support. The second and third authors would like to thank the Japan Society for the Promotion of Science for financial supports from grant-in-aids for scientific research (numbers 20730205 and 21730205), respectively. We are solely responsible for any errors and misinterpretations of this paper.

1. Introduction

Does an identical good share an identical price across geographically distinct places? Many of recent papers approach this fundamental question of the law of one price (LOP) exploiting micro-level information of retail prices observed across retail stores internationally as well as domestically. Since the seminal works by Parsley and Wei (1996) and Engel and Rogers (1996), one of the most robust findings within the literature of the LOP is a statistically significant effect of geographical distance on statistical properties of cross-regional retail price differentials. Given economic rationale provided by iceberg-type transportation costs, this robust finding suggests that transportation costs play a statistically significant role in the observed violations of the absolute LOP hindering cross-regional arbitrages of products.

The size of the distance effect that is commonly estimated in this literature, nevertheless, seems economically subtle. Regressing the absolute values of the logs of price differentials of identical products that are surveyed in two retail stores in two distinct regions on the logs of the corresponding geographical distances, many of past studies infer less than about 3 % elasticity of price differential with respect to distance.[1] This means that even when the geographical distance of two distinct cities becomes double, the price differential of a product between the cities increases at best only 3 % on average. Since the standard deviation of the absolute values of the logs of retail price differentials is typically reported around 20 % in this literature, we need the standard deviation of the logs of distances of 6.666 (=0.20/0.03) if we want to explain the whole price dispersions only by cross-regional distances. The required standard deviation of the logs of distances, however, is too large to be consistent with actual data.[2] A natural inference from this observation is that transportation costs, which are approximately measured by distances, could not be a main economic source of regional price dispersions. In this sense, a geographical distance is empirically "dead" as a prime suspect for the commonly observed violation of the LOP.

What is further puzzling is the fact that past studies of international trade unambiguously recognize that geographical distance plays economically crucial roles in determinations of bilateral trade directions and volumes. For example, Anderson and van Wincoop (2003) estimate a gravity model of bilateral trade volumes controlling for multilateral trade resistance and infer the distance elasticity of transportation costs to be around 20 % conditional on a calibrated elasticity of substitution equal to 5. Estimating a gravity model using bilateral export volume data across 183 countries, Helpman et al.(2008) find that the distance elasticity of bilateral export volumes is about 80 % once they take into account firms' selections into bilateral exports as well as firms' heterogeneity in export volumes.[3] Interestingly, their estimate suggests a 20 % distance elasticity

[1] Among a series of past studies, for example, Broda and Weinstein (2008) observe the 1.2 % distance elasticity of the absolute log price differentials within barcode-level scanner data of retail prices at retail stores across Canadian and the U.S. cities. Engel et al.(2005) find the same distance elasticity of 0.32 % with annual panel data distributed by Economic Intelligence Unit (EIU) that covers retail prices of 100 consumer goods sampled in 17 Canadian and the U.S. cities. Ceglowski (2003) reports 1.6-2.0 % estimates of the distance elasticities of 45 different products across 25 cities in Canadian provinces. Baba (2007) scrutinizes the Japanese and Korean retail price survey data and estimates less than about 3 % of the same distance elasticity after taking into account the border effect between the two countries.

[2] For instance, the standard deviation of the log of distance between two prefectural capital cities in Japan is 0.803 over all the 1081 city-pairs from 47 prefectures.

[3] In their meta analysis based on 1,051 past estimates of distance effects, Disdier and Head (2008) report the average of 0.893.

of transportation costs once we calibrate the price elasticity of demand equal to 5 as in Anderson and van Wincoop (2003). Since these studies also exploit iceberg-type transportation costs to characterize their gravity equations, the huge discrepancy in terms of the estimated size of the distance elasticity of transportation costs between the above two research agenda — the absolute LOP and the gravity model of international trade — is indeed an empirical challenge students of international economics need to explore profoundly.

In this paper, we tackle this empirical "death of distance" in regional price dispersions by making three contributions. First, this paper investigates a unique daily data set of wholesale prices of agricultural products in Japan. We follow the spirit of Parsley and Wei (1996) by using disaggregate price data within a country to avoid any potential effects of cross-country differences in tax and currency on our inference on transportation costs. Scrutinizing information of wholesale prices helps us make our estimate of transportation costs immune against local distributional costs as well as retailers' pricing strategies. More importantly, there are two outstanding characteristics of this data set: (i) we can identify the wholesale prices of an identical product at both producing and consuming regions and (ii) we can also grasp daily delivery patterns of an identical product from the former region to the latter. The first characteristic is essential for identifying transportation costs because, as discussed by Anderson and van Wincoop (2004), only when the source region of a product is identified, correct information of a trade cost could be extracted from relative prices at consuming regions to the corresponding source region. The main difficulty past studies face is in the fact that a retail price survey at retail stores rarely provides information of the source regions of a product and the market prices prevailed in these regions. Our data set, on the other hand, shows us not only in which region in Japan a variety of fruits and vegetables are produced but also at what wholesale prices these products are sold in their originated regions.[4]

Identification of the source region of a product, however, immediately leads to another fundamental question: how far a product is delivered from the source region? The second outstanding aspect of our data set empirically shows us the answer to this question. As the second contribution of this paper, we build a model to explain the observed patterns of product delivery and claim theoretically that ignoring the underlying choice of delivery might result in a serious under-bias toward our inference on the role of distance in regional price dispersions. To see this, suppose that transportation costs are unobservable and comprise two components: the one increasing proportionally in geographical distance and the other unobservable. A rise in transportation costs increases the price of the product at a consuming region and depresses the corresponding local demand for the product. Given the shape of the demand function, this fall in local product demand then tends to lower the profitability and, as a result, the probability of delivery to the consuming region from the producing region. Since the price of the product at a consuming region is observed only when a product delivery indeed occurs, an inference drawn only from information of price differentials could be subject to a sample-selection bias due to an incidental data truncation. In particular, the direction of the potential bias should be downwards because a rise in the unobservable component of transportation costs in general increases a price differential but deteriorates a probability of delivery at the same time.

[4] In a recent paper, Inanc and Zachariadis (2010) identify source regions of products reported in the Eurostat survey in several indirect ways and find around 10 % distance elasticity of price differentials in the 1990 survey. This could be indirect evidence that identification of the origin of a product is essential for inference of transportation costs.

In this paper, following Melitz (2003) and Helpman et al.(2008), we build a simple structural model of cross-regional product-delivery in which cross-regional price differentials and delivery patterns are jointly determined by the same structure of transportation costs. We then show that the degree of a sample-selection bias depends critically on two structural parameters of the model: the elasticity of transportation costs to distance and that of demand to price. Our theoretical analysis implies that drawing a correct inference on transportation costs requires us to estimate these two elasticities jointly. To do so, we propose a structural sample-selection model, which consists of the price differential and sample-selection equations, imposing nonlinear theoretical restrictions on the joint probability distribution of data. We develop a full information maximum likelihood (FIML) estimator for the empirical model. Our Monte Carlo experiments based on the model not only show us that given the price elasticity of demand, the degree of sample selection depends positively on the distance elasticity of transportation costs but also uncover two crucial facts: (i) the standard exercise of regressing price differentials on the corresponding distances provides a heavily downwards-biased estimate of the true distance elasticity of transportation costs and (ii) our FIML estimator successfully identifies the distance elasticity.

Finally, as the third contribution of this paper, we estimate our sample-selection model by FIML using the data of wholesale prices of several vegetables. The estimated sample-selection model passes two diagnosis criteria in that it does a fairly good job in replicating the actual delivery patterns of these vegetables as well as the actual data association of price differentials with distances. We find large estimates of the distance elasticity of transportation costs across all the vegetables relative to the existing estimates in the LOP literature: all of them are more than 30 % and their average is about 40 %. Given the 40 % distance elasticity of transportation costs, we need only the standard deviation of the logs of distances of 0.25 (=0.2/0.4) if we want to explain only by distances the whole part of the commonly observed standard deviation of the logs of price differentials of 20 %. The estimate of this paper, therefore, implies an economically dominant role of transportation costs in regional price dispersions. It is worth noting that this large distance elasticity does not necessarily stem from a particular characteristic of the product category of agricultural products. To prove this, we also conduct the OLS regression exercise without respecting the selection mechanism using our wholesale price data. Interestingly, we obtain the conventional range of the OLS estimate of the distance elasticity about 3 %. This provides evidence that conventional estimates of the distance elasticity could be heavily biased downwards and spuriously underestimate the role transportation costs play in regional price dispersions and LOP violations.

The organization of the rest of this paper is as follows. In the next section, we introduce our model and derive our FIML estimator based on the corresponding sample-selection model in Section 3. In Section 4, we conduct Monte Carlo experiments to check the validity of the FIML estimator. Section 5 describes our data set. After reporting the empirical results in Section 6, we conclude.

2. A simple model of cross-regional product delivery

The empirical analysis of this paper is based on the model of monopolistic competitive firms as in Melitz (2003) and Helpman et al. (2008). In this model, a country consists of I distinct regions indexed by $i = 1, 2, \cdots, I$. In each region i, the representative household consumes a continuum of agricultural products indexed by l, which takes a value between the closed unit interval, i.e.,

$l \in [0, 1]$. We assume that the representative household in each region can buy an identical set of agricultural products at the regional wholesale market and raise its utility with the Dixit-Stigliz type constant elastic function

$$u_i = \left[\int_0^1 x_i(l)^\alpha dl \right]^{1/\alpha}, 0 < \alpha < 1,$$

where $x_i(l)$ is the consumption (index) of product l in region i. This utility function implies that the elasticity of substitution across products is $\epsilon = 1/(1 - \alpha) > 1$, which is assumed to be common across all regions. The resulting region i's demand function for product l under the corresponding average price $p_i(l)$ is

$$x_i(l) = \left[\frac{p_i(l)}{p_i} \right]^{-\epsilon} x_i, \tag{1}$$

where p_i represents the the consumer price index (CPI) aggregated over products in region i

$$p_i = \left[\int_0^1 p_i(l)^{1-\epsilon} dl \right]^{1/(1-\epsilon)}, \tag{2}$$

and $x_i \equiv u_i$ indicates the indirect utility represented as the aggregate consumption level of fruits and vegetables in the corresponding region i.

We assume that each product l can be produced in all regions with an identical production technology discussed below. Producing region j of product l, then, delivers its product to the wholesale markets in the same region j as well as distinct consuming regions $i \neq j$ only if the delivery is profitable. Let $x_i(j, l)$ denote the demand of region i for product l produced in and delivered from region j. Then, the representative household in region i earns its utility from consuming product l with the following constant elastic function

$$x_i(l) = \left[\int_{j \in B_i(l)} \{\delta_i(j, l) x_i(j, l)\}^\alpha dj \right]^{1/\alpha}, 0 < \alpha < 1,$$

where $B_i(l)$ is the set of the producing regions that deliver product l to region i. This utility function specific to product l implies that the representative household in region i recognizes products indexed by l that are produced in distinct regions differently. Moreover, term $\delta_i(j, l)$ reflects the representative household's biased preference on different producing regions: the greater the term $\delta_i(j, l)$ is, the more the household in region i prefers product l produced in region j relative to those produced in other regions, *ceteris paribus*. The above CES utility function then implies region i's demand function for product l produced in region j under the price $p_i(j, l)$

$$x_i(j, l) = \left[\frac{p_i(j, l)}{p_i(l)} \right]^{-\epsilon} \delta_i(j, l)^{\epsilon - 1} x_i(l), \tag{3}$$

where $p_i(l)$ is the the aggregate price index of product l in region i

$$p_i(l) = \left[\int_{j \in B_i(l)} \{\delta_i(j, l) p_i(j, l)\}^{1-\epsilon} dj \right]^{1/(1-\epsilon)}. \tag{4}$$

As specified by Helpman et al. (2008), a producer in region j produces a unit of an agricultural product with costs minimizing a bundle of factor inputs. The marginal cost of producing product l is denoted by $c_j a(l)$, where $a(l)$ measures the number of bundles of factor inputs used per unit output of product l and c_j measures the cost of this bundle of factor inputs. Notice that $a(l)$ is product-specific, while c_j is region-specific. This means that the efficient combination of inputs for producing an identical product is common across regions, while factor input costs are different across regions.

If a producer yielding product l in region j sells its product within the same region, the delivery cost of its product to the wholesale market is $c_j a(l)$. That is to say, a producer of a region does not need to bear any transportation costs when selling its product at the wholesale market in the same region. On the other hand, if the same producer seeks to sell its product at the wholesale market in region $i \neq j$, two types of delivery costs should be borne by the producer: a fixed cost of serving at the market in region i, denoted by $c_j f_{ij}$, and an "iceberg"-type transportation cost, denoted by τ_{ij}. As in Helpman et al. (2008), we assume that $f_{jj} = 0$ for any j and $f_{ij} > 0$ for $i \neq j$, and $\tau_{jj} = 1$ for any j and $\tau_{ij} > 1$ for $i \neq j$.[5]

A producer in region j is a monopolistically competitive producer at the wholesale markets in the same region as well as the other regions to deliver. At the wholesale market in region j, the producer of product l, who faces the demand function (3), maximizes profits by charging markup price $p_j(j,l) = c_j a(l)/\alpha$. If the same producer sells at the wholesale market in region $i \neq j$, the optimal price to set, $p_i(j,l)$, is

$$p_i(j,l) = \tau_{ij} \frac{c_j a(l)}{\alpha}. \tag{5}$$

In this case, the operating profits of delivering product l to region i is

$$\begin{aligned} \pi_{ij}(l) &= (1-\alpha) \left[\frac{\tau_{ij} c_j a(l)}{\alpha} \right]^{1-\epsilon} \delta_i(j,l)^{\epsilon-1} p_i(l)^{\epsilon} x_i(l) - c_j f_{ij}, \\ &= (1-\alpha) \left[\frac{\tau_{ij} c_j}{\alpha p_i} \right]^{1-\epsilon} \theta_i(j,l)^{1-\epsilon} p_i x_i - c_j f_{ij}, \end{aligned}$$

where $\theta_i(j,l)$ is the ratio of the productivity level to the producing regional bias $a(l)/\delta_i(j,l)$. If a producer in region j sells its product l at its regional wholesale market, its monopolistic profit $\pi_{jj}(l)$ is positive because $f_{jj} = 0$ and $\tau_{jj} = 1$. However, delivering the same product to region i is profitable only if $\theta_i(j,l)$ is smaller than a threshold $\bar{\theta}_{ij}$, where $\bar{\theta}_{ij}$ is defined by the zero profit condition $\pi_{ij}(l) = 0$, or equivalently

$$(1-\alpha) \left[\frac{\tau_{ij} c_j}{\alpha p_i} \right]^{1-\epsilon} \bar{\theta}_{ij}^{1-\epsilon} p_i x_i = c_j f_{ij}. \tag{6}$$

Let $T_{ij}(l) = 1$ denote a positive delivery of product l from region j to region i, and $T_{ij}(l) = 0$ denote the zero delivery of product l from region j to region i. The above determination of the

[5] If we allow for a fixed cost of production, i.e., $f_{jj} > 0$, a producer in region j decides whether to produce product l or not depending on the corresponding zero profit condition. Appendix A, however, discusses that the sample selection due to this extensive margin of production does not result in a biased estimate of transportation costs because the decision of production of a product is independent of the transportation cost of delivering the product to other regions. We, therefore, ignore the extensive margin of production in the rest of our exercise below.

threshold (6), then, implies

$$T_{ij}(l) = \begin{cases} 1 & \text{if} \quad \theta_i(j,l) < \bar{\theta}_{ij}, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Therefore, equations (6) and (7) describe the selection mechanism of the delivery of product l produced in region j to the wholesale market in region i profitably.

Optimal price (5) implies that price differentials of an identical product between regions that are producing and consuming the product provides a precise identification of transportation cost τ_{ij}. To see this, let $q_{ij}(l)$ denote the log of the price differential of product l between producing and consuming regions j and i: $q_{ij}(l) \equiv \ln p_i(j,l) - \ln p_j(j,l)$. Then, optimal price (5) and selection mechanism (7) together yield the price differential equation

$$q_{ij}(l) = \ln \tau_{ij}, \quad \text{only if } T_{ij}(l) = 1. \tag{8}$$

Price differential equation (8) has two important empirical implications. First, transportation cost τ_{ij} can be measured from the corresponding price differential only when we can identify the prices in the producing and consuming regions. This is exactly the argument of Anderson and van Wincoop (2004) against the conventional approach to measuring transportation costs in the literature of regional and cross-country price dispersions.[6] The second implication, however, says that identifying producing and consuming regions is not enough for estimating transportation costs precisely. Equation (8) shows that there is an incidental truncation or sample section: we can observe the price differential of product l between producing and consuming regions only when the product is indeed delivered from the former region to the latter. Hence, the sample is non-randomly selected by the selection mechanism of (6) and (7). This selection mechanism indeed depends on transportation cost τ_{ij}. Therefore, transportation cost τ_{ij} in equation (8) could be inconsistently estimated unless we can take into account this sample-selection mechanism.

3. Empirical framework of identifying distance elasticity

In this section, we discuss the empirical framework for estimating transportation cost τ_{ij} that is identified by the model the last section describes. Following Helpman et al.(2008), we specify transportation cost τ_{ij} parametrically with $D_{ij}^{\gamma} \exp(\mu + u_{ij})$ where D_{ij} represents the symmetric distance between regions i and j, and $u_{ij} \sim N(0, \sigma_u^2)$ is an i.i.d. unobserved region-pair specific element of the transportation cost. Positive constant $\mu > 0$ makes it possible that the transportation cost always takes a value greater than 1 for all (i,j) pairs. The price differential (8), then, is

$$q_{ij}(l) = \mu + \gamma d_{ij} + u_{ij}, \quad \text{only if } T_{ij}(l) = 1. \tag{9}$$

The discrete choice of product l to be delivered from region j to region i is determined by threshold $\bar{\theta}_{ij}$ that is defined by zero profit condition (6). Let latent variable $Z_{ij}(l)$ denote

$$Z_{ij}(l) = \frac{(1-\alpha)\left[\frac{\tau_{ij} c_j}{\alpha p_i}\right]^{1-\epsilon} \theta_i(j,l)^{1-\epsilon} p_i x_i}{c_j f_{ij}}.$$

[6] Inanc and Zachariadis (2009) follow this argument of Anderson and van Wincoop (2004) and estimate transportation costs identifying producing countries in several indirect ways.

Product l, then, is delivered from region j to region i only if $Z_{ij}(l) > 1$. We assume that the fixed cost of delivery, f_{ij}, is stochastic due to an i.i.d. unobserved element v_{ij}. Just as in Helpman et al. (2008), we exploit a parametric specification of f_{ij}: $f_{ij} \equiv \exp(\lambda_j + \lambda_i - v_{ij})$, where $v_{ij} \sim i.i.d.N(0, \sigma_v^2)$ and is uncorrelated with u_{ij}. The log of the latent variable $Z_{ij}(l)$, $z_{ij}(l) \equiv \ln Z_{ij}(l)$, is

$$z_{ij}(l) = \beta - (\epsilon - 1)\gamma d_{ij} + \epsilon \ln p_i + \ln x_i + \xi_j + \lambda_i + \omega_l - \varrho_{ijl} + \eta_{ij}, \tag{10}$$

where $\beta \equiv \ln(1-\alpha) + (\epsilon - 1)\ln \alpha + (1-\epsilon)\mu$, $\xi_j \equiv -\epsilon \ln c_j - \lambda_j$, $\omega_l \equiv (1-\epsilon)\ln a(l)$, $\varrho_{ijl} \equiv (1-\epsilon)\ln \delta_i(j, l)$, and $\eta_{ij} \equiv (1-\epsilon)u_{ij} + v_{ij} \sim i.i.d.N(0, (1-\epsilon)^2\sigma_u^2 + \sigma_v^2)$. Selection equation (10) then implies that $T_{ij}(l) = 1$ only if $z_{ij}(l) > 0$.

Price differential equation (9) and selection equation (10) jointly reveal two critical aspects when identifying the distance elasticity of transportation costs, γ. First, estimating the distance elasticity of transportation costs only respecting price differential equation (9) might lead to an under-biased inference. To see this, taking the conditional expectation of price differential equation (9) on the observations and $T_{ij}(l) = 1$ yields

$$E[q_{ij}(l)|., T_{ij}(l) = 1] = \mu + \gamma d_{ij} + E[u_{ij}|., T_{ij}(l) = 1],$$

where . represents other observable. Notice that $E[u_{ij}|., T_{ij}(l) = 1]$ is related to the conditional expectation $\bar{\eta}_{ij} \equiv E[\eta_{ij}|., T_{ij}(l) = 1]$ by $E[u_{ij}|., T_{ij}(l) = 1] = corr(u_{ij}, \eta_{ij})\frac{\sigma_u}{\sigma_\eta}\bar{\eta}_{ij}$, where $corr(u_{ij}, \eta_{ij})$ is the correlation coefficient between u_{ij} and η_{ij} and $\sigma_\eta = (1-\epsilon)^2\sigma_u^2 + \sigma_v^2$. A consistent estimate of $\bar{\eta}_{ij}$ is obtained with the inverse Mills ratio $\hat{\bar{\eta}}_{ij}(l) = \phi[\hat{z}_{ij}(l)]/\Phi[\hat{z}_{ij}(l)]$, where $\phi(.)$ and $\Phi(.)$ are the standard normal density and cumulative distribution, respectively.[7] Therefore, we can rewrite price differential equation (9) as

$$q_{ij}(l) = \mu + \gamma d_{ij} + \beta_{u\eta}\hat{\bar{\eta}}_{ij}(l) + e_{ij}(l), \tag{11}$$

where $\beta_{u\eta} = corr(u_{ij}, \eta_{ij})\frac{\sigma_u}{\sigma_\eta}$, and $e_{ij}(l)$ is an i.i.d. error term satisfying $E[e_{ij}(l)|., T_{ij}(l) = 1] = 0$. Now, our model implies that given $\epsilon > 1$, the error term in the selection equation, η_{ij}, could be correlated negatively with that in the price differential equation, u_{ij}: $corr(u_{ij}, \eta_{ij}) < 0$.[8] Moreover, the inverse Mills ratio $\hat{\bar{\eta}}_{ij}(l)$ is increasing in distance because $\hat{\bar{\eta}}_{ij}(l)$ is a decreasing function of the predicted latent variable $\hat{z}_{ij}(l)$ that then depends negatively on distance through selection equation (10). Hence if we ignore the second term of the RHS of the above equation (11) when estimating distance elasticity γ only through the price differential equation, the resulting estimate could be biased downwards.

Second, the size of the under-bias depends crucially on the price elasticity of demand, ϵ, This is because, given an unobserved transportation cost u_{ij} and the resulting price set in the

[7] This is because

$$\begin{aligned}\bar{\eta}_{ij}(l) &= E[\eta_{ij}|., T_{ij}(l) = 1],\\ &= E[\eta_{ij}|., z_{ij}(l) > 0],\\ &= E[\eta_{ij}|., \eta_{ij} > -\hat{z}_{ij}(l)],\\ &= \frac{\phi[\hat{z}_{ij}(l)]}{\Phi[\hat{z}_{ij}(l)]}.\end{aligned}$$

[8] Because $\eta_{ij} = (1-\epsilon)u_{ij} + v_{ij}$ and u_{ij} and v_{ij} are orthogonal, $corr(\eta_{ij}, u_{ij}) = \frac{(1-\epsilon)\sigma_u}{\sqrt{(1-\epsilon)^2\sigma_u^2 + \sigma_v^2}} < 0$, given $\epsilon > 1$.

consuming region, selection equation (10) implies that a larger price elasticity leads to a smaller demand for the corresponding product sold in the consuming regional wholesale market and, as a result, lesser profitability of the delivery of the product from the producing to the consuming regions. Therefore, the under-bias due to the sample selection becomes worse with a larger price elasticity of demand. Moreover, the effect of distance on the delivery choice depends on the distance elasticity of transportation costs as well as the price elasticity of demand in a consuming region in a nonlinear way. This is because given the two elasticities, longer distance of delivery raises the price in the consumer region, reduces its demand for the product, and, as a result, depresses the profitability of delivery from the producing region. The sensitivity of the choice of delivery to distance is then nonlinearly associated with the two elasticities: the higher the distance elasticity of transportation costs is, the smaller the chance of delivery is; similarly, the higher the price elasticity of demand is, the smaller the chance of delivery is.

The above empirical implications of our model require that to identify the distance elasticity correctly, we jointly estimate the distance elasticity of transportation costs and the price elasticity of demand within a sample-selection model that consist of equations (9) and (10). For this purpose, we conduct a full information maximum likelihood (FIML) estimation of a sample-selection model on which we impose nonlinear constraints, conditional on the observations of the delivery index $\{T_{ij}(l)\}$, the price differential $\{q_{ij}(l)\}_{T_{ij}(l)=1}$, the log of distance $\{d_{ij}\}$, the average price of product l in consuming regions $\{p_i(l)\}$, and the aggregate transaction of agricultural products in consuming regions $\{x_i\}$.[9] To implement the FIML procedure, we normalize the selection equation (10) by setting the standard deviation of its error term η_{ij}, σ_η, equal to 1.[10] The normality of the distributions of the two error terms u_{ij} and v_{ij}, then, provides the following log likelihood

$$\begin{aligned}&\sum_{i,j}(1-T_{ij}(l))\ln\left[\Phi\left(-\beta+(\epsilon-1)\gamma d_{ij}-\epsilon\ln p_i-\ln x_i-b_{ij}\right)\right]\\&\quad+\sum_{i,j}T_{ij}(l)\ln\left[\Phi\left(\frac{\beta-(\epsilon-1)\gamma d_{ij}+\epsilon\ln p_i+\ln x_i+b_{ij}+\rho\sigma_u^{-1}(q_{ij}(l)-\mu-\gamma d_{ij})}{(1-\rho^2)^{\frac{1}{2}}}\right)\right]\\&\qquad+\sum_{i,j}T_{ij}(l)\ln\left[\phi\left(\frac{q_{ij}(l)-\mu-\gamma d_{ij}}{\sigma_u}\right)\right]-\sum_{i,j}T_{ij}(l)\ln\sigma_u,\end{aligned}\tag{12}$$

where $\Phi(.)$ is the standard normal cumulative distribution; $\phi(.)$ is the standard normal density; constant b_{ij} controls for regional fixed effects in selection equation (10); ρ is the correlation coefficient between u_{ij} and η_{ij}: $\rho = (1-\epsilon)\sigma_u$. Maximizing the log likelihood function with respect to the parameters γ, ϵ, b, μ, and ρ provides the corresponding FIML estimates.

4. Monte Carlo experiments with a linear economy

In this section, we conduct Monte Carlo experiments based on our model in Section 3 to understand the following two questions: (i) what bias the conventional regression exercise without

[9] We also include into our FIML estimation monthly dummies to control for seasonality and fixed effects for any unobservable producing and consuming regional effects.

[10] This standard normalization in a sample-selection model makes the correlation between u_{ij} and η_{ij} equal to $(1-\epsilon)\sigma_u$. During estimation, we further impose a restriction that the correlation coefficient $(1-\epsilon)\sigma_u$ is always less than or equal to 1 in the absolute value.

identifying producing and consumption regions and ignoring the sample-selection mechanism introduces into an estimate of transportation cost γ, and (ii) how well our FIML estimator can correct the bias.

Consider an economy that is geographically separated into 47 regions. The 47 regions are indexed by integers between 1 and 47, respectively.[11] We assume that the distance between regions i and j, D_{ij}, is equal to $100|i-j|$ with the minimum distance of 100 and the maximum of 4600. In each region, a product l is produced with productivity level $a(l)$ equal to 1 that is common across the regions. We assume that the shape of the demand function is common across the regions and characterized by the parameter α equal to 0.75. This calibration of α means that the price elasticity of demand $\epsilon = 1/(1-\alpha)$ is 4.00 and the wholesale price is 33.33 % marked up over the corresponding marginal cost. All the producing regions share the same factor cost c_j of 0.55. Each region is also characterized by the aggregate price and the aggregate real expenditure p_i and x_i, respectively, both of which we set to 20.00. For simplicity, we ignore the cross-regional variations in the productivity-regional bias ratio $\theta_i(j,l)$ by setting $\delta_i(j,l) = 1$ for all pairs of regions i and j. The fixed cost $f_{ij} = \exp(\lambda_i + \lambda_j - v_{ij})$ is specified as follows. We calibrate the sum of the producing and consuming regional fixed effects, $\lambda_i + \lambda_j$, so that, when the distance elasticity is zero, the probability of delivery of a product from a producing region to a consuming region is always equal to 0.50. The resulting fixed effect term $\lambda_i + \lambda_j$ is $(1-\alpha)\alpha^{\epsilon-1}c_j^{-\epsilon}p_i^{\epsilon}x_i$ for all (i,j) pairs. The Gaussian random component in the fixed cost, v_{ij}, has the standard deviation of $\sigma_v = 0.30$. We set the constant term of the transportation cost μ to 1.50 and allow for idiosyncratic random variations in the transportation cost setting the standard deviations of the random component of the transportation cost, σ_u, to 0.30

In our Monte Carlo experiment, we first draw 1000 sets of Gaussian random variables u_{ij} and v_{ij} independently from their distributions. We then calculate price differential $q_{ij}(l)$ and latent variable $z_{ij}(l)$ following equations (9) and (10) under one of the two hypothesized values of the distance elasticity of transportation costs, $\gamma = 0.00$, 0.15, or 0.50. In each Monte Carlo draw with each true value of the distance elasticity, we then implement four different estimations of the distance elasticity. The first one is the simple OLS regression of price differential $q_{ij}(l)$ on the log of the distance $\ln d_{ij}$ using the whole synthetic samples regardless of $T_{ij} = 0$ or 1. By construction, this OLS estimator, denoted by $\hat{\gamma}_{\text{whole}}$, is consistent and, hence, should be distributed around the hypothesized true distance elasticity. The second one is the OLS regression of the price differential $q_{ij}(l)$ on the log of the distance $\ln d_{ij}$ using only the samples that are selected with $T_{ij}(l) = 1$. This second OLS estimator, denoted by $\hat{\gamma}_{\text{OLS}}$, suffers from a sample-selection bias. Therefore, we expect to observe that the distribution of $\hat{\gamma}_{\text{OLS}}$ is biased against the true value of γ. The third estimation is with the FIML estimator we introduce in Section 3. This estimator, denoted by $\hat{\gamma}_{\text{FIML}}$, should correct potential bias due to sample selection as long as the underlying maintained assumptions are met in our Monte Carlo experiment. Finally, to explain the fourth estimator, consider the price differential between two consuming regions *without identifying producing and consuming regions*, i.e., $\tilde{p}_i(l) - \tilde{p}_k(l)$ for any two consuming regions i and k, where $\tilde{p}_i(l)$ denotes the sampled price of product l at consuming region i. The OLS estimator of the distance elasticity that is conventional in the literature of the absolute LOP, which is denoted by $\hat{\gamma}_{\text{conv}}$, then is constructed by regressing the absolute value of the price differential between consuming regions i and k, $|\tilde{p}_i(l) - \tilde{p}_k(l)|$, on the

[11] This assumption of the linear economy might be the most relevant for an island country of a long-narrow arc shape like Japan that consists of 47 prefectures.

log of the corresponding distance $\ln d_{ik}$.[12] Comparing the distribution of $\hat{\gamma}_{\text{conv}}$ with that of $\hat{\gamma}_{\text{whole}}$, we can understand the degree of bias the conventional regression exercise suffers from about the inference of the distance elasticity of transportation costs.

We first observe how the distance elasticity affects the choice of delivery. The left, middle, and right windows of Figure 1 depict the contour plots of probabilities of delivery from producing regions to consumption regions that are calculated out of 1000 Monte Carlo draws for the cases of $\gamma = 0.50$, 0.15, and 0.00, respectively. In each window, the contour lines represent sets of the producing and consuming regions that have an identical delivery probability from the former regions to the latter. The left window shows that with the large distance elasticity of $\gamma = 0.50$, product delivery is profitable only locally. This is obvious from the fact that all contour lines are parallel to the 45 degree line and equiprobability bands, which are constructed by two contour lines with the same probability, are very narrow and always include the 45 degree line. This shape of the contour plot implies that product delivery occurs only to consuming regions very close to producing regions. The middle window then exhibits that the equiprobability bands become much wider with the smaller distance elasticity of $\gamma = 0.15$. In this sense, a larger distance elasticity creates geographical clustering of products from different originated regions. This is clearer if we set the distance elasticity to zero. As displayed in the right window, the equiprobability line with the delivery probability of 0.50 are distributed over the whole window. This means that product delivery occurs with the 50 % chance even between the producing and consuming regions that are farthest apart each other.

Figure 2 depicts price differentials simulated from the model. The first, second, third rows of the figure correspond to the cases with $\gamma = 0.50$, 0.15, and 0.00, respectively. In each case, the first column reports samples that are selected by delivery choice $T_{ij} = 1$, while the second column plots the whole samples regardless of delivery choice $T_{ij} = 0$, or 1. The two windows in the first row reveal severe sample truncation under the large distance elasticity of $\gamma = 0.50$. Although the whole samples of the simulated price differentials are distributed all over the range of the log of distance and have a clear positive association with the log of distance, the underlying delivery selection mechanism is so strong that the observed samples are concentrated only on local areas with short range transportation. The association of the observed price differentials with the log of distance then becomes vague. The second and third rows prove that this sample selection turns out to be weaker when the distance elasticity becomes smaller to 0.15 and 0.00.

Figure 3 reports non-parametrically smoothed densities of the four different estimators of the distance elasticity with the Epanechnikov kernel. The first row corresponds to the case with the hypothetical value of the distance elasticity $\gamma = 0.50$; the second the case with $\gamma = 0.15$; and the third the case with $\gamma = 0.00$. The first column plots smoothed densities of the OLS estimate with the whole sample, $\hat{\gamma}_{\text{whole}}$; the second the OLS estimate with the truncated sample, $\hat{\gamma}_{\text{OLS}}$; the third the FIML, $\hat{\gamma}_{\text{FIML}}$; and the fourth the estimate of the conventional regression, $\hat{\gamma}_{\text{conv}}$. The three windows in the first column show that the OLS estimator with the whole sample, $\hat{\gamma}_{\text{whole}}$, is consistent and distributed around the underlying true value. The three windows in the second column, however,

[12] For each Monte Carlo draw, the price of product l that would be sampled at consuming region i, $\tilde{p}_i(l)$, is constructed as follows. For each consuming region i, we obtain the set of the truncated prices that are delivered from producing regions $S_i(l) = \{p_i(j,l)|j \in B_i(l)\}$. This set $S_i(l)$ includes the prices of product l that would be sampled as the representative price at consuming region i, $p_i(l)$. We uniformly draw 100 prices from this set $S_i(l)$ and take the average over them to construct $\tilde{p}_i(l)$.

uncover that the OLS estimator with the truncated sample, $\hat{\gamma}_{\text{OLS}}$, is subject to severe downwards bias. On the one hand, as displayed in the first and second rows in the second column, this estimator $\hat{\gamma}_{\text{OLS}}$ is distributed far left from the corresponding true value when the true distance elasticity is either 0.50 or 0.15. On the other hand, as shown in the third row of the second column, if the true distance elasticity is zero, the OLS estimator with the truncated sample is consistent and distributed around the true value of $\gamma = 0.00$. Therefore, the positive distance elasticity generates the sample truncation that causes the OLS estimates to be biased downwards. The three windows in the third column reports the smoothed densities of the FIML estimator for the three true values of the distance elasticity. These windows clearly reveal that the FIML estimator is consistent and distributed around the underlying true value. Finally, the three windows in the fourth column plot the smoothed densities of the OLS estimator of the conventional regression, $\hat{\gamma}_{\text{conv}}$. The most outstanding fact from these windows is that the conventional estimator performs the worst. In the first and second rows for the cases of $\gamma = 0.50$ and 0.15, the conventional estimator $\hat{\gamma}_{\text{conv}}$ is distributed with the means of 0.019 and 0.003, respectively, and even far left from the corresponding distributions of the OLS estimator $\hat{\gamma}_{\text{OLS}}$. This clearly shows us that the conventional regression exercise without identifying producing regions indeed suffers from the worst under-bias toward the inference on the distance elasticity of transportation costs among all the other estimators.

The Monte Carlo experiment of this section, therefore, confirms the necessity of identifying producing and consuming regions and taking into account the sample-selection mechanism for drawing a correct inference on the distance elasticity of transportation costs. The FIML estimator can correctly identify the true values of the distance elasticity with synthetic data generated from our structural model.

5. Data and descriptive statistics

In this paper, we investigate a unique daily data set of the wholesale prices of agricultural products in Japan: the Daily Wholesale Market Information of Fresh Vegetables and Fruits ("*Seikabutsu Hinmokubetsu Shikyo Joho*").[13] Appendix provides the detailed description of the data set. All contents in the data set are surveyed by the Ministry of Agriculture, Forestry, and Fishery for almost all transactions at 55 wholesale markets officially opened and operated in the 47 prefectures in Japan on a daily basis. This daily market survey covers the wholesale prices of 120 different vegetables and fruits that are actually traded in the 55 wholesale markets. Each agricultural fresh product is further categorized by different varieties, sizes, grades, as well as producing prefectures. Hence, for example, the data set reports the wholesale prices at 6 different wholesale markets of "Dansyaku (Irish Cobbler equivalent)" variety of potato of size "L" with grade "Syu (excellent)" produced in "Hokkaido" prefecture on September 7, 2007. This high degree of categorization is ideal for our purpose of approaching the absolute LOP rigorously and inferring transportation costs precisely because the law requires to identify identical goods as its theoretical premise at the first place, as Broda and Weinstein (2009) seek in their barcode data. This daily market survey has been recorded since 1976. In this paper, however, we select the 2007 survey in which there are reported market transactions on 274 market opening days.[14]

[13] The data set is distributed by the Center of Fresh Food Market Information Service ("*Zenkoku Seisen Syokuryohin Ryutsu Joho Senta*") with the URL: http://www2s.biglobe.ne.jp/ fains/index.html.

[14] The choice of 2007 is arbitrary. We are extending our exercise to other years.

Price differential $q_{ij}(l)$ is constructed by subtracting the wholesale price in the producing prefecture, $p_j(j,l)$, from that in the consuming prefecture, $p_i(j,l)$. By construction, $q_{jj}(l) = 0$ for any producing prefecture j.[15] We assign the value of 1 to delivery index $T_{ij}(l)$ for any (i,j) combination with observed $q_{ij}(l)$.[16] The geographical distance between prefectural combination (i,j) is approximated by that between the corresponding prefectural head offices in the prefectural capital cities. The data of distance is provided by the Geographical Survey Institute (GSI) of the Government of Japan. Taking the logarithm of the geographical distance yields variable d_{ij}.[17] We cannot obtain daily data of the aggregate price over all agricultural products implied by selection equation (10), p_i. Variable p_i, hence, is proxied by the monthly data of the retail price of the corresponding vegetable that is reported in the Retail Price Survey ("*Kouri Bukka Tokei Cyosa*") the Ministry of Internal Affairs and Communication conducts. Selection equation (10) also implies the aggregate consumption level over all agricultural products in consuming region i, x_i, as a factor determining whether to delivery. To construct data for variable x_i, we take the sum of daily volumes of 24 different vegetables traded at the wholesale market of consuming region i. Moreover, to control for daily variations in producing and consuming prefectures, we include into selection equation (10) daily temperature data in both of the two prefectures that are reported by the Japan Meteorological Agency.[18] This inclusion of the regional temperatures as determinants of delivery comes from our prior belief that the temperatures in producing and consuming regions are important factors for productions of and demands for agricultural products.

We focus our exercise on selected vegetables: cabbage, carrot, Chinese cabbage (c-cabbage, hereafter), lettuce, shiitake-mushroom (s-mushroom, hereafter), spinach, potato, and welsh onion. Table 1 summarizes several descriptive statistics for these vegetables observed in the 2007 survey. The table shows that each vegetable is highly categorized by product variety, sizes, and grades. In this paper, we consider products having different source regions as different products even when they are of the same product category. The number of distinct product entries, then, is quite large; 1,207 for cabbage; 1,186 for carrot; 1,001 for c-cabbage; 903 for lettuce; 1,423 for potato; 909 for s-mushroom; 551 for spinach; and 1,115 for welsh onion, respectively.

For each product entry l, we count the numbers of delivery $T_{ij}(l) = 1$ and non-delivery $T_{ij} = 0$ only for the dates on which the product entry l is indeed traded at the wholesale market in the producing prefecture j. The seventh row of the table reports that the total number of both delivery and non-delivery cases all over the product entries is beyond two hundred thousands for each vegetable. This is the number of observations for our FIML estimation. Out of the total number of delivery and non-delivery cases, the number of delivery cases is relatively small, as exhibited in the eighth row of the table: it is around ten thousands for each vegetable. The seventh and eighth rows of the table, therefore, imply that product delivery from a producing prefecture to consuming prefectures is quite limited. More informatively, the ninth row shows that the average distance from producing regions to consuming regions over all delivery and non-delivery cases

[15] For some products, we cannot find the wholesale price in the producing prefecture, $p_j(j,l)$, although we can observe those prices in the consuming prefectures, $p_i(j,l)$. In this case, because we cannot construct the price differential between producing and consuming prefectures, we drop these products from our investigation.

[16] We also assign the value of 1 to $T_{jj}(l)$ whenever we can observe the wholesale price in the producing prefecture $p_j(j,l)$. We consider this case that the corresponding product l is delivered from the producer to the wholesale market in the producing prefecture. We attach the minimum distance of 10km to these samples with $T_{jj}(l) = 1$ to avoid taking the logarithm of zero distance.

[17] The data is publicly available in the GSI website, http://www.gsi.go.jp/KOKUJYOHO/kenchokan.html.

[18] We download daily temperature data from the website: http://www.data.jma.go.jp/obd/stats/etrn/index.php.

$T_{ij} = 1$ or 0 is almost the same across the vegetables and about 6.00 in the logarithmic term (or 403.428 km). The tenth row, on the other hand, conveys that the average distance over all delivery cases is much shorter depending on the vegetables with the minimum number of 2.69 (14.77km) for s-mushroom and the maximum of 4.35 (77.55km) for potato.[19] The critical aspect of our data set the ninth and tenth rows uncover, hence, is the fact that product delivery is localized and concentrated around the corresponding producing prefecture.

Figure 4 also confirms graphically the locality of product delivery. Each window of the figure depicts as a contour plot the frequencies of product delivery from producing prefectures to consuming prefectures that are calculated over all product entries on all traded dates. The horizontal axis represents producing prefectures and the vertical axis consuming prefectures. The order of prefectures reflects the geographical positions of the prefectures from the most north prefecture, Hokkaido, to the most south one, Okinawa. Therefore, two prefectures that are indexed by close integers are indeed geographically close to each other. Then, the brighter the blue contour line is, the higher the probability of product delivery is. Hence, we can expect that contour lines should be concentrated on the 45 degree line if product delivery is completely localized around the producing prefectures. We observe the following three things in the figure. First, each vegetable has several dominant producing prefectures that have vertically concentrated contour lines. This means that these main producing prefectures deliver their products to not only nearby prefectures but also other remote prefectures. Second, however, the frequencies of product delivery of the main producing prefectures are decreasing in distance. Even dominant producers do not deliver their products to consuming prefectures farthest away.[20] Third, the contour lines for other minor producing prefectures are concentrated on the 45 degree line. The product delivery of these relatively minor producing prefectures, thus, is highly localized.

The locality of product delivery that Table 1 and Figure 4 unmask together brings us two important implications. First, as observed by Broda and Weinstein (2008) in their barcode data of retail products, agricultural products in our data set are segmented and clustered geographically. Even in the same vegetable category, products that are sold in two distinct regions far away one another come from distinct producing origins and the corresponding prices might be affected by idiosyncratic regional factors of producing regions. Variations in price differentials across consuming regions that are generated by these idiosyncratic factors of distinct origins cannot be attributed to transportation costs. Hence, given the observed high degree of regional product clustering, it is crucial to scrutinize regional price differentials of a product that shares the same product origin in order to infer the role transportation costs play in the absolute LOP deviation. Second, drawing an inference on transportation costs only from observed price differentials might be subject to a serious sample-selection bias, as we repeatedly claim in this paper.

The averages of the observed log price differentials are reported on the first row of the lower panel of Table 1. The positive numbers of the row imply that prices in consuming prefectures are on average higher by between 0.3 % and 8.1 % than those of the producing prefectures. This observation of higher prices in final destinations than those in the origins are suggestive for an important role of transportation costs in price differentials, as predicted by equation (9). The

[19] This short average distance of delivery reflects the fact that products are almost always delivered to the wholesale markets of their producing prefectures. In this case of $T_{jj}(l) = 1$, as noted by footnote 14, we assign the minimum distance of 10.00km to the observations.

[20] Exception is observed in producing prefecture 1, Hokkaido, in the cases of carrot and potato.

corresponding standard deviations of the observed price differentials, which are displayed on the second row of the lower panel, are quite large and around 20 %. This observed degree of regional price dispersions is quite common in the literature of the LOP (e.g., Crucini et al. 2005, and Broda and Weinstein 2008). Our data set of wholesale prices of agricultural products shows the almost same degree of the LOP violation as observed in these past studies, even after we identify products that share the same producing regions. We also conduct an OLS regression of the observed price differentials on the corresponding log distances and constant for each vegetable. The resulting OLS estimates of the coefficient on the log distance, $\hat{\gamma}_{\text{OLS}}$, are shown in the third row of the lower panel of Table 1, which are accompanied by the corresponding standard errors. All the estimates are positive and statistically significant with values between the minimum of 0.007 and the maximum of 0.051. This range of the estimated distance elasticity of price differential is consistent with the estimates observed in the past studies using different data sets such as Engel et al. (2005), Broda and Weinstein (2008), and Inanc and Zachariadis (2009).

What is striking in the lower panel of Table 1 is that our price differential data of agricultural products are characterized by many important data aspects that are frequently emphasized by the past studies of the LOP using their distinct micro-level data of retail prices. This fact implies that the inference of the distance elasticity of transportation costs by estimating the sample-selection model by FIML does not necessarily result from data characteristics of regional price differentials of particular agricultural products. In the next section, we report the results of the FIML estimation of our sample-selection model.

6. Results

Table 2 summarizes the results of the FIML estimation of the sample-selection model. The first row of the table shows that the elasticity of transportation costs with respect to distance, $\hat{\gamma}_{\text{FIML}}$, is estimated positive and statistically significant for each vegetable. The outstanding fact this row tells us is the large size of all the FIML estimates: the average (over the eight vegetables) of the estimated distance elasticities is 0.399 with the minimum of 0.301 for cabbage and the maximum of 0.522 for shiitake mushroom. According to price differential equation (9), the price differential of a product between consuming and producing regions rises by about 40 % in response to the 100 % stretch in delivery distance when we ignore selection mechanism (10). Given the small size of the OLS estimates of the distance elasticities for the eight vegetables, which are reported between 0.7 % and 5.1% in Table 2, this large size of the FIML estimates implies that the OLS estimates are biased downwards seriously due to the underlying data truncation.

As discussed in section 3, the strength of the observed under bias tightly connects with the price elasticity of demand, ϵ. As reported in the second row, the elasticity of demand with respect to price, ϵ, is estimated both precisely and sensibly: the average of the estimated values of the price elasticity of demand is 3.744 over the eight vegetables. Therefore, the demand for each vegetable in consuming prefectures is reasonably responsive to a change in the wholesale price of the product. Combining with the large estimate of the distance elasticity of transportation costs, the estimated price elasticity of demand implies that the probability of product delivery from producing to consuming prefectures depends negatively as well as sensitively on delivery distance. Indeed, the estimates of the correlation coefficient between the unobserved disturbances of price differential equation (9) and selection equation (10), ρ, provide evidence that sample-selection bias does matter.

As displayed in the third row of the table, the correlation coefficient is estimated negative with high statistical significance for each vegetable: the average of the correlation coefficients over the vegetables is -0.779 with the minimum of -0.543 for shiitake mushroom and the maximum of -0.866 for carrot. This highly negative correlation between the unobserved disturbances in the two equations is the fundamental source for the under bias in the OLS estimate of the distance elasticity in the price differential equation, as shown in equation (11).

In summary, our FIML estimates of the sample-selection model reveal dual roles geographical distance plays in regional price differentials. Distance creates a large price gap between consuming and producing regions. At the same time, distance significantly affects choice of delivery from the latter to the former regions. As a result, price differentials are not randomly sampled and, especially, their observations are concentrated on local areas surrounding producing regions. This concentration of the observations within relatively short distance conceals the actual size of the underlying distance elasticity of transportation costs making the OLS estimate biased downwards.

Model validation through diagnosis checks

The above FIML estimates of the three structural parameters depend on the identification provided by our structural sample-selection model. Therefore, the relevance of the estimates relies on the empirical validity of our model. As model validations, we conduct diagnosis checks of our model with respect to two important aspects of the actual data: the pattern of product delivery and the association of price differential with delivery distance.

If our sample-selection model is reliable, it should match the pattern of product delivery, $T_{ij}(l)$, that is actually observed in our data. To check the ability of our model to mimic the delivery pattern in the data, we calculate the percent correctly predicted (PCPs) measures for $T_{ij}(l) = 0$ or 1. To construct the PCPs, we calculate the predicted conditional probabilities of $T_{ij} = 0$ and $T_{ij} = 1$ on the observable, $\hat{P}(T_{ij} = 0|.)$ and $\hat{P}(T_{ij} = 1|.)$, respectively.[21] Then if $\hat{P}(T_{ij} = 0|.) > 0.5$, we predict $T_{ij} = 0$. Similarly, if $\hat{P}(T_{ij} = 1|.) > 0.5$, we predict $T_{ij} = 1$. Then, the PCP for $T_{ij}(l) = 0$ (or 1) is calculated as the percentage of the total number of the observations of $T_{ij}(l) = 0$ (or 1) that are accompanied by $\hat{P}(T_{ij} = 0|.) > 0.5$ (or $\hat{P}(T_{ij} = 1|.) > 0.5$). The PCP for either $T_{ij}(l) = 0$ or 1 is simply derived as a weighted average of the two PCPs.

The results of the PCPs are summarized in the first, second, and third rows of the lower panel of Table 2. As shown in the first row, our sample-selection model yields high PCPs around 0.99 for either $T_{ij}(l) = 0$ or 1 for all the vegetables. This means that the model is fairly successful in replicating the observed pattern of product delivery overall. In particular, as implied by the PCPs reported in the second and third rows of the lower panel, the model's ability to replicate no delivery choice $T_{ij}(l) = 0$ is better than that to replicate delivery choice $T_{ij}(l) = 1$. On the one hand, the high PCPs for no delivery choice around 0.99, which are reported in the second row of

[21] The conditional probabilities, $\hat{P}(T_{ij} = 0|.)$ and $\hat{P}(T_{ij} = 1|.)$, are calculated as

$$\hat{P}(T_{ij} = 0|.) = \Phi\left(-\hat{\beta} + (\hat{\epsilon} - 1)\hat{\gamma} d_{ij} - \hat{\epsilon} \ln p_i - \ln x_i - \hat{b}_{ij}\right),$$

and

$$\hat{P}(T_{ij} = 1|.) = \Phi\left(\frac{\hat{\beta} - (\hat{\epsilon} - 1)\hat{\gamma} d_{ij} + \hat{\epsilon} \ln p_i + \ln x_i + \hat{b}_{ij} + \hat{\rho}\hat{\sigma}_u^{-1}(q_{ij}(l) - \hat{\mu} - \hat{\gamma} d_{ij})}{(1 - \hat{\rho}^2)^{\frac{1}{2}}}\right),$$

respectively.

the lower panel, suggest the model's almost perfect predictive ability of no delivery choice. On the other hand, the PCPs for delivery choice, which are reported in the third row, are lower with the cross-vegetable average of 0.800. The model does a good job in predicting the delivery choice $T_{ij}(l) = 1$ especially for some vegetables such as s-mushroom, spinach, and welsh onion.[22] We confirm through this diagnosis criterion that our model's predictive performance for the pattern of product delivery is remarkable.

The second diagnosis criterion is data association of price differential with distance. As observed in Table 1, the OLS regression of the observed price differential on delivery distance yields the estimate of the distance elasticity $\hat{\gamma}_{\text{OLS}}$ around 4 % on average. The question we ask here is if our sample-selection model predicts this size of the OLS estimate or not.

To do this diagnosis check, we derive the prediction of the estimated model on price differential following equation (11). Each window of Figure 5 plots the resulting predicted price differentials (blue dots) as well as their data counterparts (gray crosses) against the corresponding log distances for each vegetable. Observe that in each window the blue dots are distributed inside the cloud made of the gray crosses. This means that our model successfully predicts the data association of price differential with distance for each vegetable, although the actual data show us a much sparse joint distribution between price differential and distance. The fourth row of the lower panel of Table 2 reports the OLS estimate $\hat{\gamma}_{\text{OLS}}$ of regressing the predicted price differentials on the corresponding distances. For comparison, we also repeat in the last row the OLS estimates with the actual data that are reported in Table 1 too. The model's predictions on the OLS estimates are close to but slightly larger than their actual data counterparts: the cross-vegetable average of the predicted OLS estimate is 0.078 whereas that with the actual data is 0.033. It is important, however, to remember that the distance elasticity of transportation costs of our model is estimated 0.399 by FIML. What is striking is that the sample-selection model with such a large distance elasticity of transportation costs indeed mimics such a small size of the OLS estimate. In this sense, we conclude that our model successfully passes the second diagnosis check, although we fully understand that there is still an unexplained gap between the model's prediction and the actual data with respect to the observed joint distribution of price differential and distance. The question is then what our model miss to fill this gap. We leave this important question as our future research task.

7. Conclusion

As claimed by Anderson and van Wincoop (2004) in their introduction, the "death of distance" is indeed exaggerated even in regional price dispersions. In this paper, we try to revive and rejuvenate transportation costs, which are measured by geographical distance, as a potential source of absolute LOP violations. In so doing, we identify producing regions and take into account sample selectivity due to the underlying choice of product delivery from producing to consuming regions in our unique data of daily wholesale prices of agricultural products in Japan. After estimating our structural sample-selection model by FIML using the data of price differentials and delivery patterns, we find that the estimated distance elasticities of transportation costs are so large

[22] The main reason for the model's slightly lower predictive performance for carrot and potato is simple. As observed in Figure 4, the main producing prefecture of these two vegetables, Hokkaido, delivers its products to all other prefecture regardless of delivery distance. This data aspect is hard to explain by our simple structural model.

among vegetables that we can fill the reported huge gap between the two fields of international economics — international finance and empirical trade — in terms of inferences on distance effects.

[*Conclusion to be continued*]

References

Anderson, J.E., van Wincoop, E., 2003, "Gravity with gravitas: a solution to the border puzzles," *American Economic Review* 93, 170 – 192.

Anderson, J.E., van Wincoop, E., 2004, "Trade costs," *Journa of Economic Literature* 42, 691–751.

Baba, C., 2007, "Price dispersion across and within countries: the case of Japan and Korea," *Journa of the Japanese and International Economies* 21, 237 – 259.

Broda, C., Weinstein, D.E., 2008, "Understanding international price differences using barcode data," *NBER working paper* 14017.

Crucini, M. J., Telmer, C.I., Zachariadis, M., 2005, "Understanding European real exchange rates," *American Economic Review* 95, 724 – 738.

Crucini, M. J., Telmer, C.I., Zachariadis, M., 2005, "Price dispersions: the role of borders, distance, and location," mimeo.

Ceglowski, J., 2003, "The law of one price: intranational evidence for Canada," *Canadian Journal of Economics* 36, 373 – 400.

Disdier, A.-C., Head, K., 2008, "The puzzling persistence of the distance effect on bilateral trade," *Review of Economics and Statistics* 90, 37 – 48.

Engel, C., Rogers, J.H., 1996, "How wide is the border?" *American Economic Review* 86, 1112 – 1125.

Engel, C., Rogers, J.H., Wang, S.-Y., 2005, "Revisiting the boarder: an assessment of the law of one price using very disaggregate consumer price data," in Driver, R., Sinclair, P., Thoenissen, C., eds., *Exchange Rates, Capital Flows and Policy*, Routledge, NY.

Helpman, E., Meitz, M., Rubinstein, Y., 2008, "Estimating trade flows: trading partners and trading volumes," *The Quarterly Journal of Economics* 123, 441 – 487.

Inanc, O., Zachariadis, M., 2010, "The importance of trade costs in deviations from the law of one price: estimates based on the direction of trade," *Economic Inquiry*, forthcoming.

Melitz, M., 2003, "The impact of trade on intra-industry reallocations and aggregate industry productivity," *Econometrica* 71, 1695 – 1725.

Parsley, D. C., Wei, S.-J., 1996, "Convergence to the law of one price without trade barriers or currency fluctuations," *The Quarterly Journal of Economics* 114, 1211 – 1236.

Wooldridge, J.M., 2002, *Econometric Analysis of Cross Section and Panel Data* MIT Press, Cambridge.

Appendix A. Does sample selection of production result in a biased estimate of γ?

Let $Y_j(l)$ denote the indicator function that takes the value of one if region j produces a fruit or vegetable l and delivers the product to its wholesale market, and the value of zero otherwise. Notice that when region j does not produce product l, we cannot observe the price differential $q_{ij}(l)$, i.e.,

$$q_{ij}(l) = \gamma d_{ij} + u_{ij}, \quad \text{only if } T_{ij}(l) = 1 \text{ and } Y_j(l) = 1.$$

Suppose that a farmer in region j has to pay a fixed cost $f_{jj} > 0$ when they decide to produce a fruit or vegetable. The profit of producing product l and delivering it to the regional wholesale market then is

$$\pi_{jj} = (1-\alpha)\left(\frac{c_j}{\alpha p_j}\right)^{1-\epsilon} \theta_j(j,l)^{1-\epsilon} y_j - c_j f_{jj},$$

and the resulting zero profit condition is

$$(1-\alpha)\left(\frac{c_j}{\alpha p_j}\right)^{1-\epsilon} \tilde{\theta}_j^{1-\epsilon} y_j = c_j f_{jj},$$

where $\tilde{\theta}_j$ is the threshold of no production. Now define a variable $W_j(l)$ such as

$$W_j(l) = \frac{(1-\alpha)\left(\frac{c_j}{\alpha p_j}\right)^{1-\epsilon} \theta_j(j,l)^{1-\epsilon} y_j}{c_j f_{jj}}.$$

The threshold $\tilde{\theta}_j$ means that region j produces product l only if $W_j(l) > 1$ or $w_j(l) \equiv \ln W_j(l) > 0$. As in the case of f_{ij}, assume that the fixed cost f_{jj} is stochastic: $f_{jj} = \exp(\lambda_j - v_j)$, where $v_j \sim i.i.d.N(0, \sigma_v^2)$ and is uncorrelated with u_{ij}. The latent variable $w_j(l)$ then is

$$w_j(l) = \gamma_0 + (\epsilon - 1)\ln p_j + \ln y_j + \xi_j + \omega_l - \mu_{jjl} + v_j.$$

This is the selection equation of the production of product l in region j. Ignoring this sample selection does not result in a biased estimate of γ in the price differential equation. This is because v_j is uncorrelated with u_{ij}. Consider the conditional equation $E[q_{ij}(l)|d_{ij}, Y_j(l) = 1]$:

$$\begin{aligned} E[q_{ij}(l)|d_{ij}, Y_j(l) = 1] &= E[\gamma d_{ij} + u_{ij}|d_{ij}, Y_j(l) = 1], \\ &= \gamma d_{ij}, \end{aligned}$$

because

$$\begin{aligned} E[u_{ij}|d_{ij}, Y_j(l) = 1] &= E[u_{ij}|d_{ij}, w_j(l) > 0], \\ &= E[u_{ij}|d_{ij}, p_j, x_j, \cdots, v_j], \\ &= 0. \end{aligned}$$

This result means that we obtain an unbiased estimate of γ, selecting the samples with $Y_j(l) = 1$ and estimating the sample-selection model by the FIML procedure.

Appendix B. Data description

The data file contains information on name of product, market prices, name of production cite, name of market place, and product characteristics. The price reported has three forms: the highest price, the mode price, and the lowest price. Most markets record all three prices, but several markets report only the highest and the lowest prices or only the mode price. Thus, we construct our price variable by averaging these price variables. We use the mode price when only the mode price is available. The transaction unit of each product is also reported. To obtain same unit for each product, we divide the price by the number of unit.

We need to control for product characteristics to examine prices between production cite and market place. Thus, we construct same category product by using product characteristics and production cite. The product characteristics are: brand name, size of products, and grade of products. The size is coded by categorical variables, such as large, medium, and small. The grade is also measured by the categorical variables, such as A, B or superior. For example, spinach is classified as grade A under the following conditions: it is of one type and no mixture of types affects the appearance; it is clean, trimmed, and free from decay and damages by insects (Source: Guideline document of Yamanashi prefecture). Otherwise, it is ranked as B. Because prices depend on detailed characteristics, we take each combination of characteristics to have the same product.

With respect to central markets, the coverage of vegetables traded through central wholesale markets is substantial in Japan. While nowadays large supermarket chains can directly purchase agricultural products from producers, the share of domestic products covered by central wholesale markets is more than 90 percent in 2006. While there are growing imports of agricultural products, the overall share (not only domestic, but also imported products) of vegetables traded is still 75 percent (Source: Ministry of Agriculture, Forestry, and Fishery). Thus, our data enable us to capture not a particular channel of distribution cost, but representative transportation costs.

Table 1 : Description of vegetable data

	Cabbage	Carrot	C-Cabbage	Lettuce	Potato	S-Mushroom	Spinach	Welsh Onion
No. of varieties	3	10	4	7	10	1	4	11
No. of size categories	63	62	50	71	50	74	17	103
No. of grade categories	34	66	50	46	93	55	85	58
No. of producing prefectures	47	46	46	43	47	44	47	46
No. of wholesale markets	47	47	47	47	47	47	47	47
No. of distinct product entries	1,207	1,186	1,001	903	1,423	909	551	1,115
No. of $T_{ij}(l) = 0$ or 1	369,343	198,129	241,871	239,703	264,280	476,919	466,347	547,272
No. of $T_{ij}(l) = 1$	15,841	8,395	10,803	11,565	10,921	11,845	15,977	14,874
Ave. log distance over $T_{ij}(l) = 0$ or 1	5.942	6.031	5.942	5.990	6.227	5.935	5.925	5.948
Ave. log distance of $T_{ij}(l) = 1$	3.707	4.008	4.009	3.950	4.351	2.693	3.255	2.943
Ave. log price differential $q_{ij}(l)$	0.039	0.076	0.065	0.026	0.081	0.003	0.029	0.016
SD. log price differential $q_{ij}(l)$	0.167	0.285	0.227	0.267	0.265	0.127	0.216	0.178
$\hat{\gamma}_{\text{OLS}}$	0.033	0.051	0.042	0.022	0.037	0.007	0.044	0.033
(s.e.)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)

Note 1: $\hat{\gamma}_{\text{OLS}}$ represents the OLS estimate of the coefficient γ in the regression specification $q_{ij}(l) = \mu + \gamma d_{ij} + u_{ij}(l)$ where μ is constat and $u_{ij}(l)$ is an OLS disturbance. Note that $q_{ij}(l)$ is the price differential between consuming and producing regions i and j. "(s.e.)" reports the corresponding standard error.

Table 2 : FIML estimation of the sample-selection model

	Cabbage	Carrot	C-Cabbage	Lettuce	Potato	S-Mushroom	Spinach	Welsh onion
$\hat{\gamma}_{\mathrm{FIML}}$	0.301	0.362	0.412	0.426	0.349	0.522	0.433	0.384
(s.e.)	(0.002)	(0.004)	(0.003)	(0.004)	(0.003)	(0.004)	(0.003)	(0.003)
$\hat{\epsilon}$	4.258	3.082	3.702	3.081	3.338	4.572	3.739	4.186
(s.e.)	(0.018)	(0.017)	(0.018)	(0.014)	(0.015)	(0.028)	(0.014)	(0.018)
$\hat{\rho}$	-0.847	-0.866	-0.826	-0.863	-0.771	-0.543	-0.835	-0.844
(s.e.)	(0.002)	(0.003)	(0.004)	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)
log likelihood	-26486.776	-20688.309	-17725.484	-25224.822	-27287.880	-5636.887	-25461.829	-20522.294
No. of observations	369,343	198,129	241,871	239,703	264,280	476,919	466,347	547,272
Diagnosis checks								
PCP for $T_{ij}(l) = 0$ or 1	0.987	0.988	0.987	0.986	0.984	0.995	0.992	0.995
PCP for $T_{ij}(l) = 0$	0.996	0.998	0.994	0.995	0.999	0.999	0.996	0.998
PCP for $T_{ij}(l) = 1$	0.800	0.782	0.817	0.807	0.630	0.863	0.838	0.861
$\hat{\gamma}_{\mathrm{OLS}}$ with predicted $q_{ij}(l)$	0.079	0.111	0.077	0.091	0.061	0.024	0.100	0.084
$\hat{\gamma}_{\mathrm{OLS}}$ with actual $q_{ij}(l)$	0.033	0.051	0.042	0.022	0.037	0.007	0.044	0.033

Note 1: The log likelihood of the FIML estimation is given by equation (11). Each estimation includes monthly dummies, consuming prefectural dummies, and producing prefectural dummies both in price differential and selection equations (9) and (10).

Note 2: "Pcp" represents the "percent correctly predicted."

Figure 1: Simulated probabilities of delivery

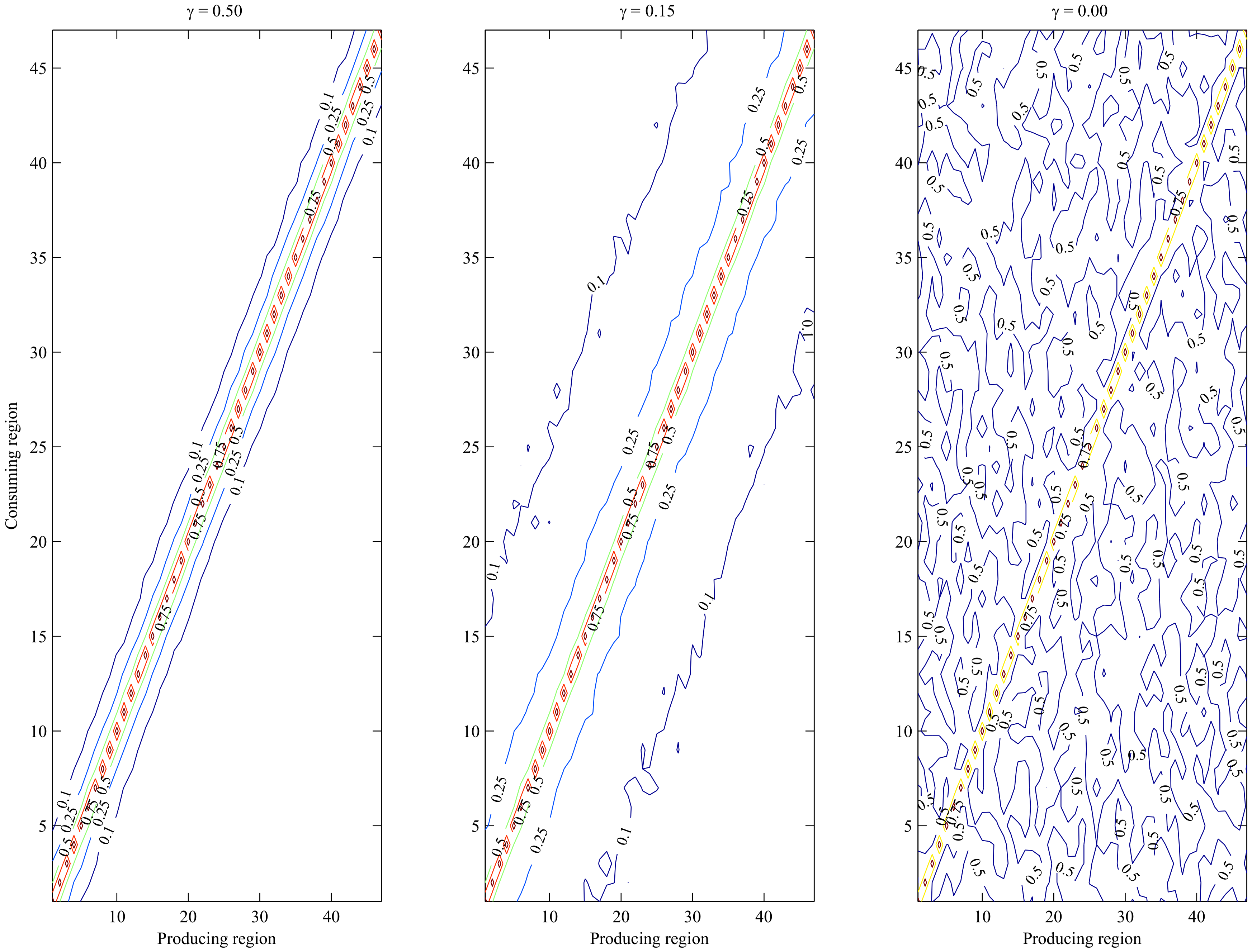


Figure 2: Simulated price differentials

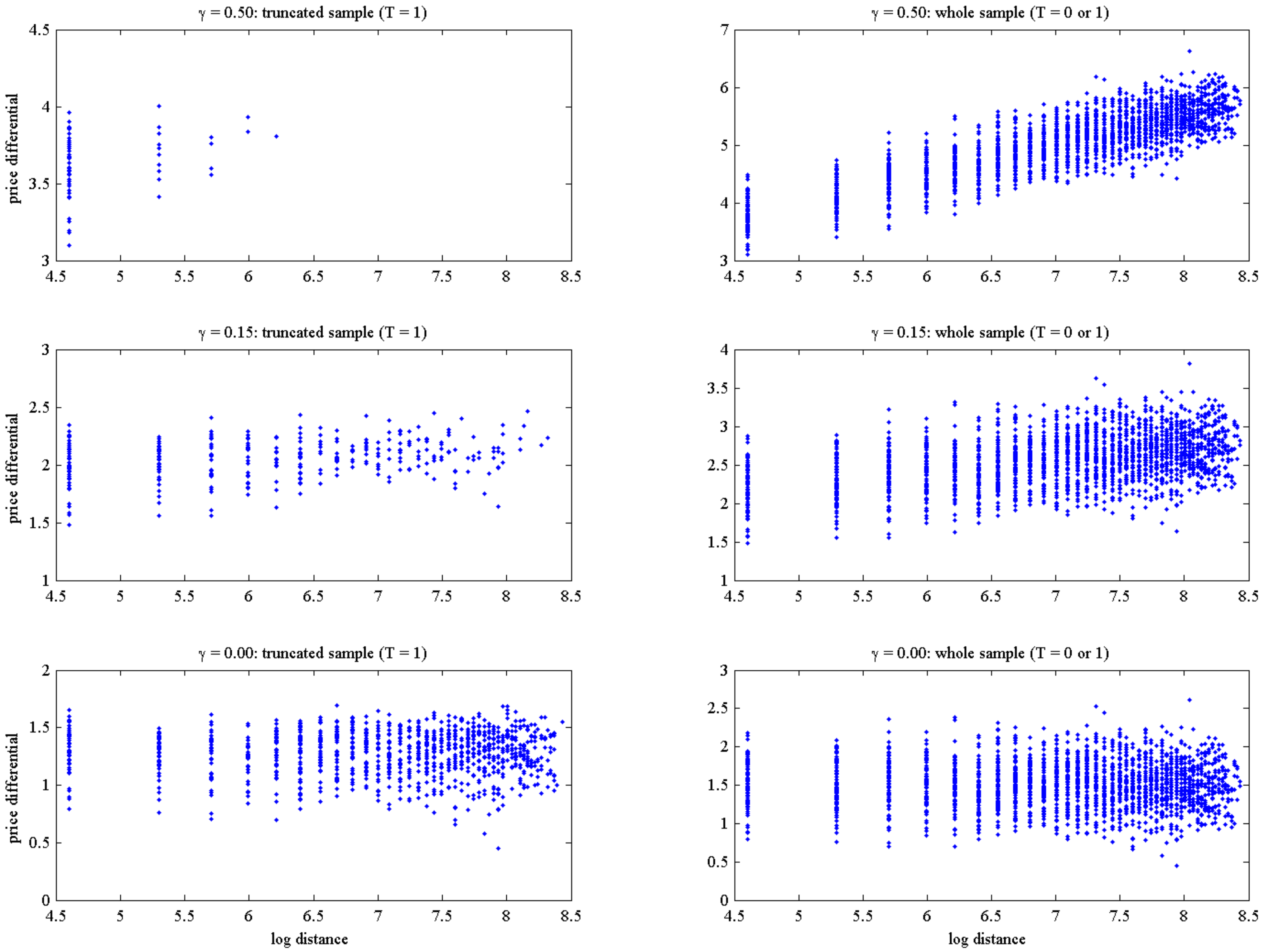


Figure 3: Kernel smoothed densities of estimators of the distance elasticity

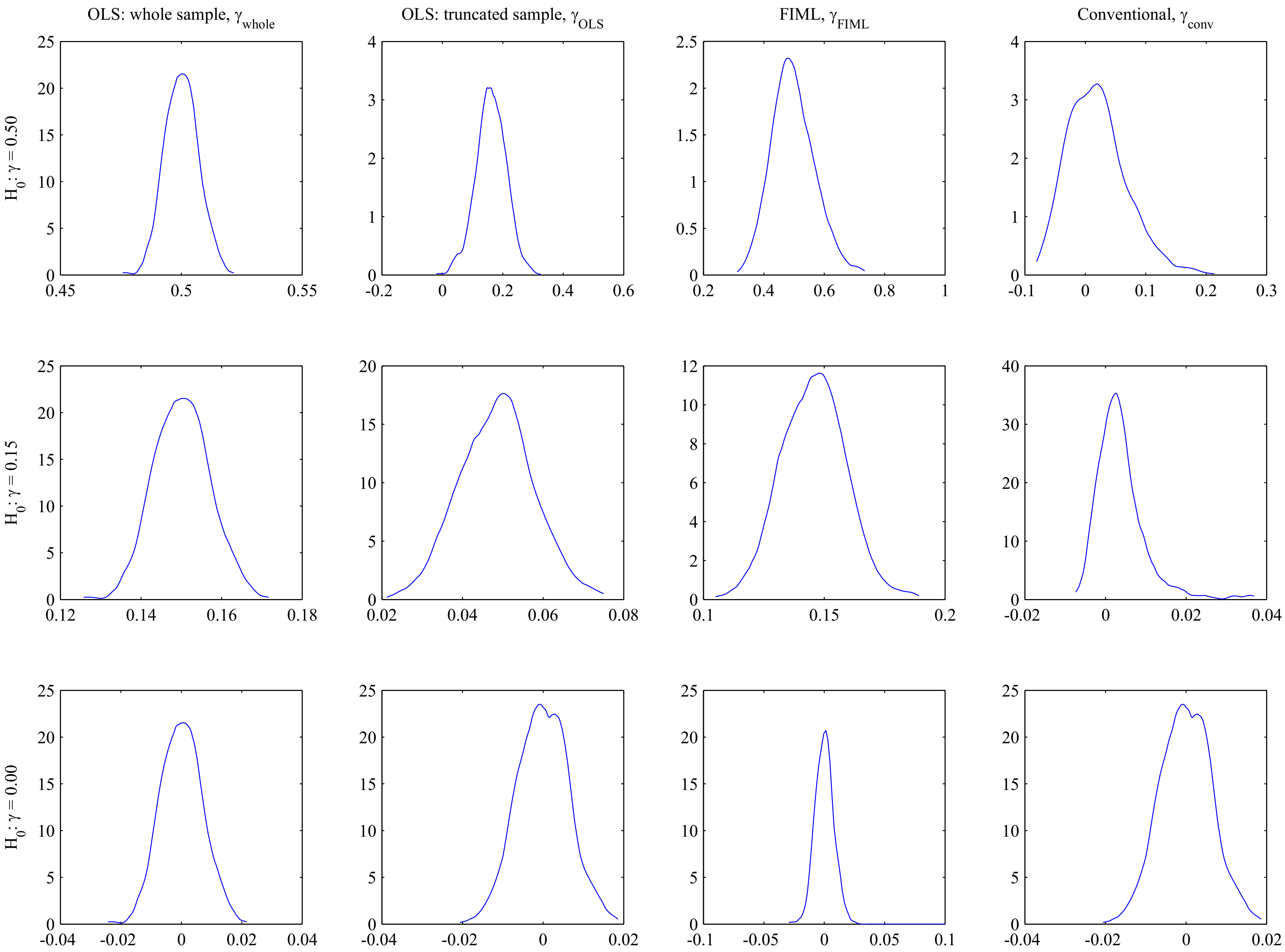
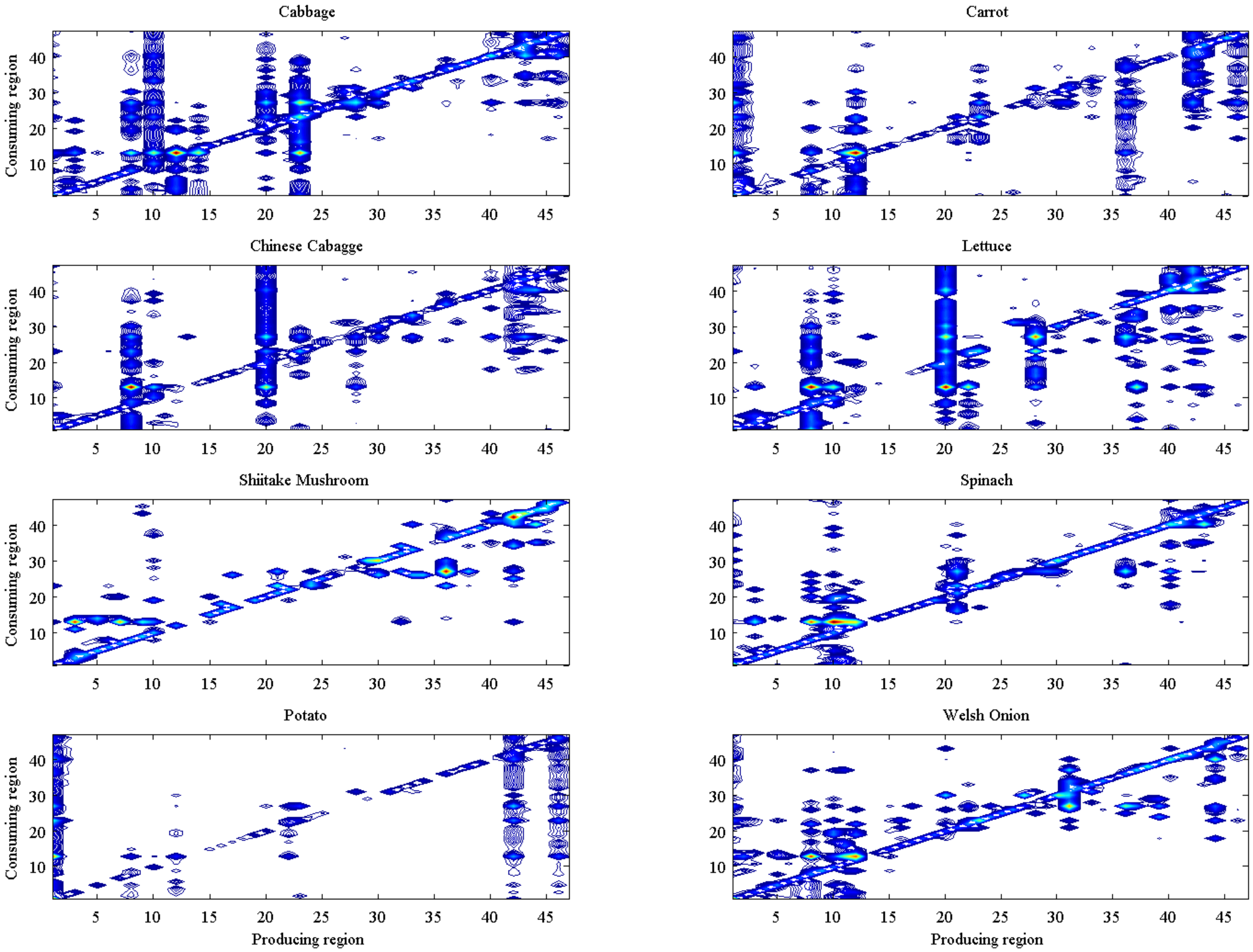


Figure 4: Data probabilities of delivery



Cabbage

Carrot

Chinese Cabagge

Lettuce

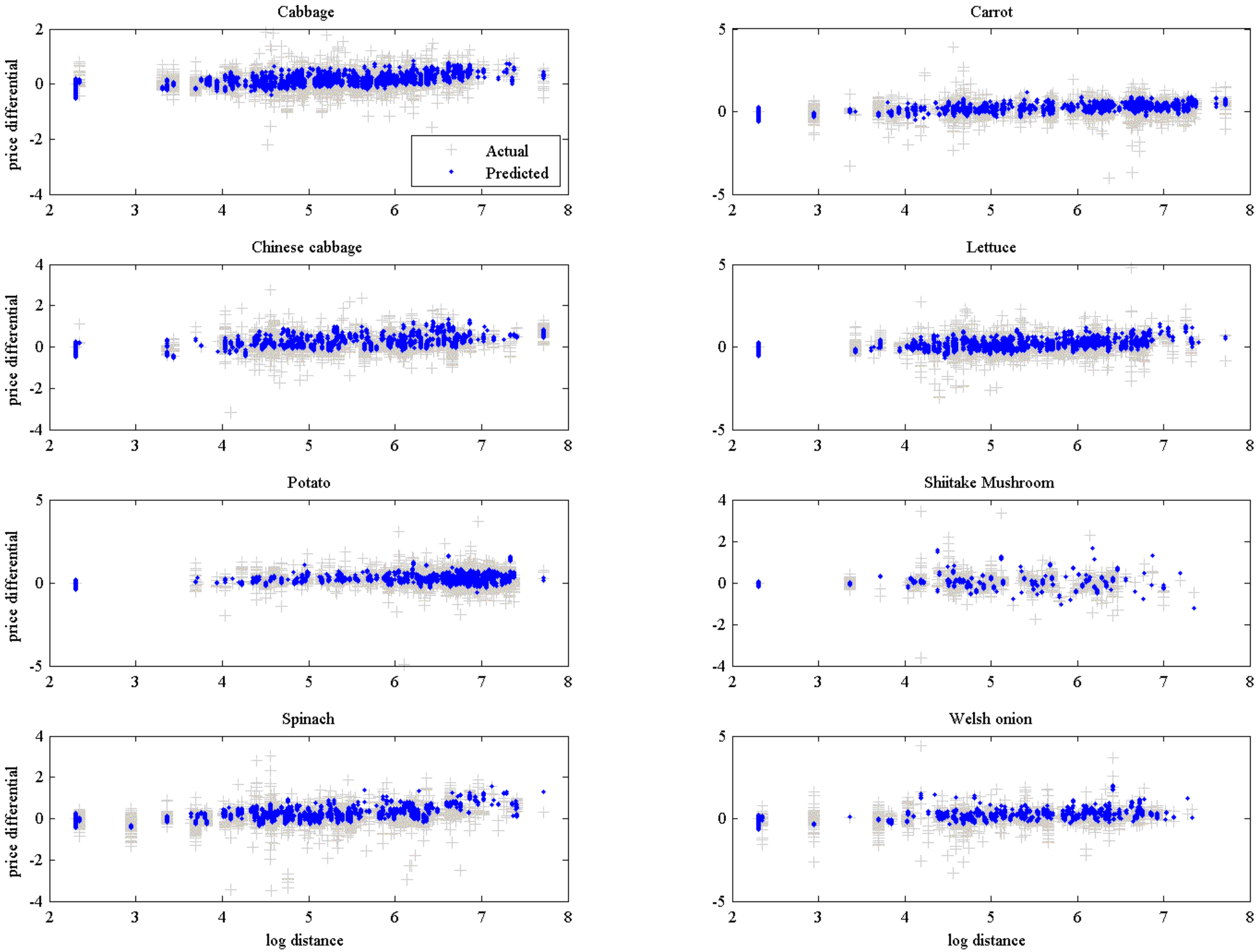
Shiitake Mushroom

Spinach

Potato

Welsh Onion

Figure 5: Predicted price differentials


Cabbage
Carrot
Chinese cabbage
Lettuce
Potato
Shiitake Mushroom
Spinach
Welsh onion
price differential
log distance
Actual
Predicted