Exploiting Big Data in Logistics Risk Assessment via Bayesian Nonparametrics

Yan Shang, David Dunson, Jing-Sheng Song

Exploiting Big Data in Logistics Risk Assessment via Bayesian Nonparametrics

In cargo logistics, a key performance measure is transport risk, defined as the deviation of the actual arrival time from the planned arrival time. Neither earliness nor tardiness is desirable for customer and freight forwarders. In this paper, we investigate ways to assess and forecast transport risks using a half-year of air cargo data, provided by a leading forwarder on 1336 routes served by 20 airlines. Interestingly, our preliminary data analysis shows a strong multimodal feature in the transport risks, driven by unobserved events, such as cargo missing flights. To accommodate this feature, we introduce a Bayesian nonparametric model – the probit stick-breaking process (PSBP) mixture model – for flexible estimation of the conditional (i.e., state-dependent) density function of transport risk. We demonstrate that using simpler methods, such as OLS linear regression, can lead to misleading inferences. Our model provides a tool for the forwarder to offer customized price and service quotes. It can also generate baseline airline performance to enable fair supplier evaluation. Furthermore, the method allows us to separate recurrent risks from disruption risks. This is important, because hedging strategies for these two kinds of risks are often drastically different.

Key words: Bayesian statistics, big data, disruptions and risks, empirical, international air cargo logistics, nonparametric, probit stick-breaking mixture model

1. Introduction

Global trade has grown considerably in recent decades; many companies now have overseas facilities and supply chain partners. International cargo logistics management thus plays an increasingly important role. Air transport delivers goods, that are time-sensitive, expensive, perishable or used in just-in-time supply networks, at competitive prices to customers worldwide. Indeed, air cargo transports approximately 35% of world trade by value (IATA 2014b) and is forecast by Boeing to grow at an average 4.7% annual rate in the next two decades. However, attention paid to this industry is surprisingly little: air cargo industry '.. has remained the poor cousin to the more glamorous passenger side of the business (passenger air transport industry)' (Morrell 2011).

The consequences of this neglect are significant as the service level of cargo transport has become firms' big concern. In cargo logistics, a key (service) performance measure is *transport risk* (or delivery reliability), defined as the deviation of the actual arrival time from the planned arrival time,

$$\text{transport risk} = \text{actual arrival time} - \text{planned arrival time}.$$

Neither earliness nor tardiness is desirable. While tardiness causes delay in production and product/service delivery to all downstream customers, earliness incurs additional storage and handling costs. Extreme risks, such as more than 48 hour delays or more than 24 hours earliness, is defined as *(transport) disruption risks*, because they severely impact the operations of the customers and the freight forwarders. To distinguish disruption risks from the routine deviations within a day, we refer to the latter as *recurrent risks*. According to a 2011 PRTM survey, 69% of companies named improving delivery performance as their top supply chain management strategy. In a 2010 report of Infosys, "carrier delays and non-performance on delivery" is ranked as the leading risk in the logistics industry. Furthermore, in a 2014 survey conducted by International Air Transport Association (IATA) to major freight forwarders and their customers, low reliability is perceived as the second most important factor (next to transportation cost).

In this paper, we study the transport risks of international air cargo based on a half-year of air cargo data between 2012 and 2013, provided by a leading forwarder on 1336 routes served by 20

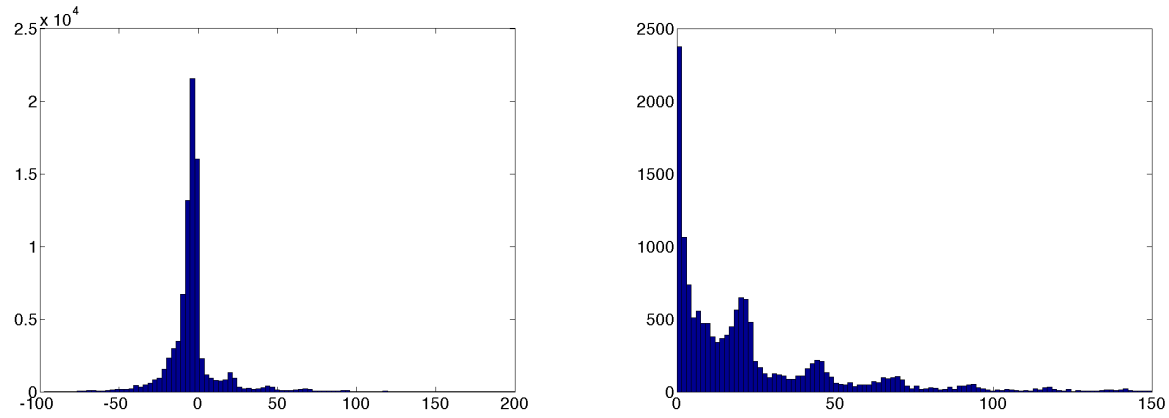

Figure 1 Histograms of transport risk (hours)

airlines. Using a Bayesian nonparametric (BNP) model – the Probit stick-breaking (PSBP) mixture model – we obtain accurate estimates of transport risk distributions and disruption risk probabilities. Our model provides a tool for the forwarder to offer customized price and service quotes. It can also generate baseline airline performance to enable fair supplier evaluation.

We make several contributions to the Operations Management (OM) literature as outlined below.

Empirical Air Cargo Transport Risk Distribution

Our work appears to be the first empirical study of global air logistics in the supply chain literature. One interesting phenomenon observed from the data is that the distribution of transport risk, conditional on predictors (i.e., independent variables including airline, route, shipping time, cargo weight etc), is a *multimodal distribution*, as shown in Figure 1. The left side of Figure 1 is the empirical distribution of transport risks of all shipments observed in the data (almost 90 thousand shipments), which, clearly, is a non-symmetric, long-tail distribution with several bumps at the distribution's positive part. To better observe the bumps, we only plot the data that falls in the range (0, 150) on the right side of Figure 1. Here, we can see clearly that big bumps concentrate around days (at 24 hours, 48 hours, and 72 hours, etc.) and small bumps concentrate between days. These systematic peaks are largely due to the fact that a cargo that failed to be loaded onto its scheduled flight was loaded onto a flight on the same route later. The scheduled gap between flights, which depends heavily on the route, for example, is usually around 24 hours for international flights

and 4 – 6 hours for domestic flights. The time gaps between scheduled flights thus transfer to the gaps between different peaks in the conditional distribution of transport risk to form a multimodal distribution; see §3 for more detail.

Previous empirical studies primarily focus on domestic passenger flight arrival or departure delays; see Deshpande and Arikan (2012) for a review. (Note that delay or lateness is the positive part of transport risk, because earliness is usually not a concern for passenger flights.) Most of this literature assumes the delays follow unimodal distributions. Under this assumption, most works, such as Shumsky (1995) and Mueller and Chatterji (2002), adopted the classic ordinary least square linear regression (OLS) for delay estimation. However, our above multimodal observation indicates that OLS is unsuitable for the air cargo transport risk assessment and prediction. This is because OLS is built on the assumption that the distribution of a dependent variable, conditional on other predictors, is unimodal (most often Normal distribution). Our work is close to Tu et al. (2008), in which the authors used a mixture model to estimate flight departure delay nonparametrically. However, they only analyzed the data from one airport by one airline. Hence, we need to develop new methodologies for our big data as described below.

BNP Model and Conditional Distribution Function

Our second contribution is methodological. To accommodate the multimodal feature in the empirical transport risk distribution, we introduce a state-of-the-art Bayesian statistics tool – the BNP mixture model. To the best of our knowledge, no prior work has used related techniques in empirical OM, which so far predominantly applies frequentist statistics, such as OLS and maximum likelihood estimation (MLE), see, e.g., Deshpande and Arikan (2012), Li et al. (2014) and the references therein.

Bayesian statistics has experienced rapid development in the past two decades accelerated by ever-increasing computational power. Among these tools, BNP mixture models have become popular in the last several years, with applications in fields as diverse as finance, econometrics, genetics, and medicine (refer to Rodriguez and Dunson (2011) for references therein). A nonparametric mixture model can be expressed as follows: in the case where we are interested in estimating a single

distribution from an independent and identically distributed (*i.i.d.*) sample $y_1, \cdots, y_n$, observations arise from a convolution

$$y_j \sim \int k\left(\cdot \mid \boldsymbol{\psi}\right) G\left(\mathrm{d}\boldsymbol{\psi}\right)$$

where $k\left(\cdot \mid \boldsymbol{\psi}\right)$ is a given parametric kernel indexed by $\boldsymbol{\psi}$ (we use bold symbol to indicate vector), and G is a mixing distribution assigned a discrete form

$$G\left(\boldsymbol{\psi}\right) = \sum_{l=1}^{L} \omega_l \delta_{\boldsymbol{\psi}_l}, \text{ where } \sum_{l=1}^{L} \omega_l = 1 \text{ and } \omega_l \geq 0, \, \forall l = 1, \cdots, L$$

and L can be finite or infinite. For example, assuming that G follows a Dirichlet process (DP) prior leads to the well known Dirichlet process mixture (DPM) model (Escobar and West 1995).

For our application, we adopt a specific BNP model – the PSBP mixture model, which was formally developed in Rodriguez and Dunson (2011). This method is known for its flexibility, generality, consistency under week regularity conditions (Pati et al. 2013) and (importantly) computational tractability. Rodriguez et al. (2009) used this technique to create a nonparametric factor model to study genetic factors predictive of DNA damage and repair. Chung and Dunson (2009) applied this tool to develop a nonparametric variable selection framework. Our model is designed to capture the transport risk distribution characteristics in all ranges, covering both recurrent and disruption risks.

Particularly, we focus on modeling the conditional distribution of transport risks, within the PSBP framework. Modeling the conditional distribution allows us to investigate the relationship between transport risks and potential predictors, including airline, route, shipping time, cargo weight etc, based on which we can further explore ways to improve transport reliability. We will explain this in more details in §3.1.

To demonstrate the value of PSBP, we compare our transportation risk estimation with that obtained from OLS. We show that the two methods deliver dramatically different results, see §3.5 and §4. For instance, OLS fails to capture the critical roles airlines play in transport service levels. More importantly, the OLS predictions underestimate disruption risks, which can result in insufficient risk management strategies.

Data-Driven Risk Assessment Tool

Our method suggests a powerful and general tool to help supply chain risk assessment, a topic that has not received the attention it deserves. A recent McKinsey & Co. Global Survey of Business Executives shows that "nearly one-quarter of firms say their company doesn't have formal risk assessment". As articulated in Van Mieghem (2011), managing risk through operations requires 4 steps to be executed and updated recurrently: 1. identification of hazards; 2. risk assessment; 3. tactical risk decisions; 4. implement strategic risk mitigation or hedging. Among the four steps, step 1 is more experience and context based, which typically involves information from anecdotal records or long experience with the specific business processes. Step 4 is more action-based, requiring detailed organizational design and information systems to carry out the hedging strategies developed. These two steps may not need quantitative methods. Steps 2 and 3, on the other hand, require rigorous analysis and quantification, and therefore call for analytical research. While most of the supply chain risk management literature focuses on the third step, which involves developing strategies for reducing the probabilities of negative events and/or their consequences should they occur, this paper focuses on step 2 – risk assessment.

Risk assessment consists estimations of two components: (a) risk likelihood, i.e., "the probability that an adverse event or hazard will occur" and (b) risk impact, i.e., "the consequences of the adverse event" (Van Mieghem 2011). The long-term expected risk is the integration of these two parts. Though scarce, we noticed a distinguished work by Kleindorfer et al. (2003) on assessing risk impact (part (b)) of catastrophic chemical accidents using data collected by the Environmental Protection Agency. Kleindorfer and Saad (2005) presented a conceptual framework for risk assessment and risk mitigation for supply chains facing disruptions. Different from these studies, our work focuses on using statistical methods to accurately estimate the risk likelihood (part (a)), which calls for more advanced scientific computation and analysis tools. Our study shows that a careful risk assessment is critical to developing tailored services for customers (i.e., shippers) of different types and selecting service suppliers.

The transport risk studied in this paper resembles the random yield/capacity risks in manufacturing studied by many authors; see, e.g., Federgruen and Yang (2009), Wang et al. (2010). Also, the transportation disruption risk is an important type or component of random supply disruption risks considered by Song and Zipkin (1996), Tomlin (2006), etc. While most of these authors assume a particular risk distribution, such as a Bernoulli distribution for disruptions, the Bayesian PSBP mixture model introduced here can be used to generate empirical random yield distributions and disruption probabilities, when data are available.

The reminder of the paper is organized as follows: in §2, we give a brief introduction of the air cargo logistics industry and its challenges, the data we used for this study and the research questions we ask. In §3, we describe exploratory analysis to lead to formal model selection, and we introduce the PSBP mixture model and the algorithm for posterior Gibbs sampling. In §4 we explain the results. In §5 we propose several applications of our model to design more efficient operational strategies. In §6, we conclude the paper and discuss future directions. Appendix A contains data summary statistics and estimation results.

2. Industry Background, Data Source, and Research Questions

Though a crucial part of global operations, the air cargo industry is less known to the public because it operates behind the scenes. For this reason, in order to understand our model and analysis, it is necessity to provide a brief background of the industry. We use *Cargo 2000* as our data source.

2.1. Service Chain Structure

Typically, an air cargo transport involves four parties: *shippers* (e.g., manufacturers), *freight forwarders* (*forwarders* in short), *carriers* (i.e., airlines) and *consignees* (e.g., downstream manufacturers or distributors); see Figure 2 for an illustration. A shipper initiates a transaction by providing the forwarder company with "(1) origin/destination; (2) collection/delivery date; (3) shipment details (cargo pieces, weight and volume); (4) shipper/consignee information; (5) product/shipping service required"(IATA 2014a). Following their route map, the forwarder picks up cargoes from the shipper at the required time, consolidating cargoes sharing the same route if possible, and then sends cargoes to the selected airline at an origin airport. The airline takes charge of cargoes until arriving at

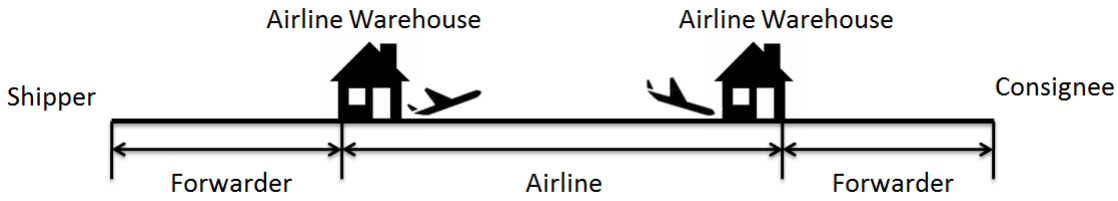

Figure 2 **Cargo flow**

the destination airport. An airline might use a direct flight or 2 – 3 connecting flights based on the route map. The forwarder accepts cargoes at the destination, and delivers them to consignees.

To simplify terms, we refer to both the shipper and the consignee as the "customers". Customers use forwarders in 90% of air cargo shipments. A forwarder is a service provider for its customers, while it in turn uses airlines as service providers. Upon receiving a shipping request, a forwarder sends a booking request to several airlines, choosing the most economic one that satisfies the agreed upon timetable. Large forwarders typically reserve a certain percentage (e.g., 30%) of the total space on most airlines, including passenger and cargo airlines.

2.2. *Cargo 2000* (C2K) Standards

To compete against integrators (service providers who arrange door-to-door transportation by combining mode(s) of transportation, such as DHL, UPS etc.), *Cargo 2000* (C2K) was founded by a group of leading airlines and freight forwarder companies, "IATA Interest Group", in 1997. This initiative was designed to enable industry-wide participants to "provide reliable and timely delivery shipments through the entire air transport supply chain" (IATA 2014a). Specifically, they developed a system of shipment planning and performance monitoring for air cargo which allows proactive and realtime event processing and control. Currently C2K is composed of more than 80 major airlines, forwarders, ground-handling agents, etc. C2K Quality Management System is implemented with two different scopes: Airport-to-Airport (A2A) and Door-to-Door (D2D). In this paper, we focus on the A2A level shipments due to data constraints. The following describes how C2K is used to create a shipping plan, and how airlines and forwarders monitor, control, intervene and repair each shipment in real-time.

2.2.1. Plan After a carrier has confirmed requested capacity on planned flights, it creates an A2A route map (RMP) and shares it with the forwarder. A RMP describes the path the freight shipment follows, including flight information as well as milestones and the latest-by time for the fulfillment of the milestones along the transport chain. If a customer agrees on the plan, the RMP is set alive. Otherwise, modifications will be made until agreement is achieved. Essentially, each route map is a combination of a station profile and milestones. Station profile, which contains information on the duration for completion of each process step, are kept by forwarders and carriers. The milestones are defined by the C2K MOP.

2.2.2. Monitor, Control, Intervene and Repair After a route map is issued, the shipping process is monitored against this map. The completion of every milestone triggers updates on both the airline's and forwarder's IT systems. Any deviation from the plan triggers an alarm, which allows for corrections to be taken by the responsible party in order to bring the shipment back on schedule. If necessary, a new RMP is made for the remaining transport steps. Meanwhile, an exception record is entered into the system recording the necessary information such as time, location, and reasons.

2.2.3. Report At the end of the shipment process, a report, including whether or not the delivery promise was kept and which party was accountable for the failure, is generated. This allows the customers to directly compare the performance of their C2K enabled forwarders, carriers and logistics providers.

2.3. Forwarder's Frustration and Our Objectives

Even with current systems, the service level remains unsatisfying. As a result, forwarders risk loosing customers even though forwarders have no direct control of A2A, which is the most uncertain part of shipping. Questions for the *forwarder* to solve include: (1) how to predict transport risks so as to prepare for risks and inform customers in advance and (2) how to improve transport reliability in each route by selecting the best supplier? We aim to help address these questions. Suppose a customer comes to the forwarder with a fixed route (origin-destination), time of shipping, weight and volume of cargo. We aim to provide the forwarder with a distribution of transport risk conditional

Table 1 Potential predictors

demand variables	
route (r)	an origin-destination airport pair combination (captures all the fixed effects on a particular route).
month (m)	month when the shipping is finished
cargo weight (wgt)	total weight of the cargo (kilograms)
cargo number-of-pieces (pcs)	total number of pieces of the cargo (unit load)
decision variables	
airline (a)	the airline transported the cargo
number of legs (leg)	number of connecting flights taken to arrival at destination
planned duration (dur)	total time (days) planned to take to finish the transport
initial deviation (dev_{start})	deviation (days) between actual and planned check-in time at airline origin warehouse

on demand variables (route, month, cargo weight/volume) and decision variables (airline, number of flight legs, planned duration, initial deviation time) with 95% uncertainty interval. See Table 1 for descriptions of these variables. Based on this information, an optimal route can be chosen to match the customer's cost/utility function, providing different options to different customers. Next, we elaborate how the above mentioned demand and decision variables affect the transport risk.

2.3.1. Effect of Demand Variables

1. Route: service level differs dramatically across routes depending on (a) supply-demand of air transport service and (b) congestion level and infrastructure at visited airports. We use a route-level effect to absorb all these factors.

2. Month: demand (e.g., holiday shipping) and weather (e.g., winter snow) both have a seasonal trend, which results in different perceived air cargo transport service levels in different months. We used the month when the transport is completed as the predictor; since shipments only take 1.7 days

to finish on average, essentially identical results would be achieved using the month of transport start.

3. Cargo weight and volume: each flight has a maximum weight and volume (cargo volume is approximated by cargo pieces in this paper). Larger cargoes may be more likely to fail to be loaded onto the scheduled flight due to (1) airlines overselling capacities and (2) changes of currently available capacity, such as more luggage from passengers. However, larger cargoes are usually more valuable, thus may have higher transport priority. Our analysis can help reveal which factor is more dominant.

2.3.2. Effect of Decision Variables

1. Airline: transportation delays are expected to vary substantially across airlines due to a wide variety of factors, and hence we added (1) the interaction of airline and route and (2) the interaction of airline and number of legs into the model.

2. Number of legs: number of legs increases the probability for a cargo to miss connecting flights, so is a strong predictor of transport risk. The number of legs is included in the models as a predictor, corresponding to a decision variable.

3. Planned duration: even conditional on route, airline and number of legs, planned duration differs greatly. This reflects cushions added to the shipping time.

4. Initial deviation: if the cargo is sent to the airline earlier than scheduled, it can be loaded onto an earlier flight and vice versa.

2.3.3. Other Potential Predictors There are other factors, such as price and weather, that may also affect the risk distribution, but are not available in our data. Our model indirectly captures these effects through allowing the distribution of risk to vary flexibly with the demand and decision variables mentioned above.

2.4. Data and Summary Statistics

Our data contains a leading freight forwarder company's C2K standard airfreight shipments from October 2012 to April 2013. The data contains real-time milestone updates, similar to the data

shown in Table A.1, and route maps for each shipment. The last route map before the shipment is used to measure risk. After cleaning, the data includes 86,149 shipments on 1336 routes operated by 20 airlines. Freights are shipped from 58 countries to 95 countries. In Appendix §A.1 are summary statistics. In sum, we observe that: (i) European airlines, such as Lufthansa and KLM, play a significant role in the data; (ii) More than 50% of shipments are transported on routes served by more than 1 airline. For example, around 30% shipments are on routes served both by direct flight and 2-leg service.; (iii) There are more than 50% of shipments transported on routes where services of different legs are available. These confirm the need for a careful assessment of the impact of different choices, which can lead to a higher utility if service levels vary significantly.

2.4.1. Exception Records Exception codes are meant to facilitate (1) finding root causes of delays and (2) identifying parties accountable for failures. Unfortunately, as confirmed by the company as well as our data, exception codes are not helpful in these regards. Less than 8% of delays are assigned exception codes, with only 10% of delays of more than 1 day coded. In addition, codes are ambiguous, with the most frequently appearing code being "COCNR", denoting the carrier hasn't received the cargo. Hence, we do not use exception data in our analysis.

3. Model

In this section, we explain the model in detail. §3.1 provides the motivation for estimating the conditional distribution of transport risk and the advantage of using PSBP mixture model. The model can be decomposed into two parts: mixture weight and mixture kernel, detailed in §3.2 and §3.3, respectively. The Bayesian posterior sampling algorithm to estimate unknown parameters in weights and kernels is presented in §3.4. We also discuss model selection, and provide comparisons with OLS in §3.5. Due to space limitation, the label switching move we use to improve sampling efficiency, the choice of Bayesian priors for parameters and model implementation are not included in this paper (upon request to the author).

3.1. Conditional Risk Distribution

The multimodal feature is not only present at the aggregate data level, see Figure 1, but also at the granular level, such as each route or route-airline level. The histograms in Figure 3 are the

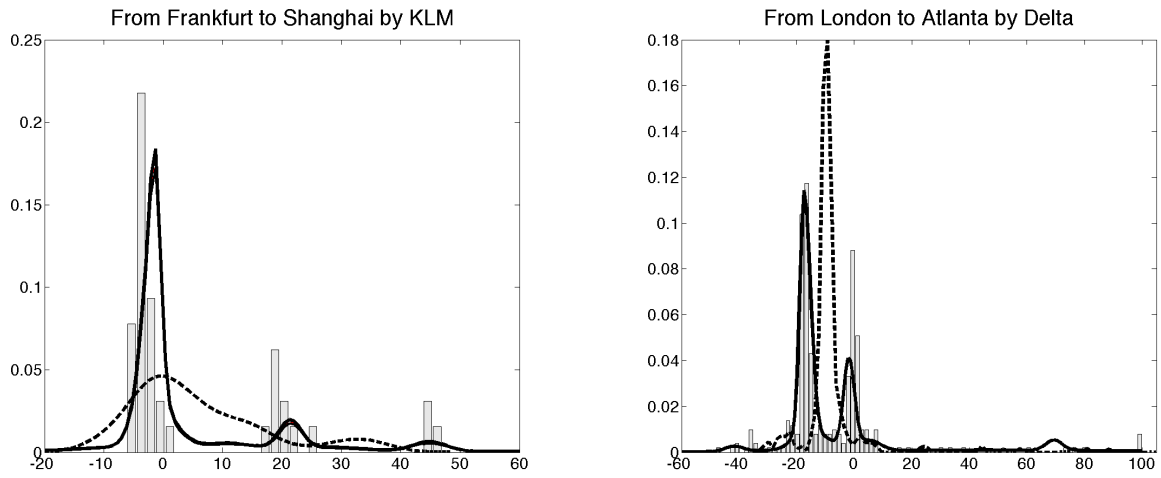

Figure 3 Sample routes

data empirical distributions on two sample routes served by two airlines. In order to make accurate predictions and inferences based on such data, the first step is to choose a model flexible enough to fit the data well. Usual choices of models for multimodal data rely on mixtures, e.g., mixtures of Normal kernels, which are known to provide an accurate approximation to any unknown density.

We cannot rely on simple mixture models, as we are investigating the distribution of transport risks conditional on demand and decision variables, including both categorical and continuous predictors. This leads to a problem of ***conditional distribution estimation***. One stream of literature on flexible conditional distribution estimation uses frequentist methods. Fan et al. (1996) proposed a double-kernel local linear approach, and related frequentist methods have been considered by Hall et al. (1999) and Hyndman and Yao (2002) among others. The other popular choice is a BNP mixture model. Muller et al. (1996) proposed a Bayesian approach to nonlinear regression, in which the authors modeled the joint distribution of dependent variable and independent variables using a Dirichlet process mixture (DPM) of Normals (Lo 1984, Escobar and West 1995). This type of approach relies on inducing a model for the conditional distribution of the response through a joint model for the response and predictors. Although such joint models are provably flexible, in practice they can have clear disadvantages relative to models that directly target the conditional response distribution without needing to model the high-dimensional nuisance parameter corresponding to

the joint density of the predictors. Such disadvantages include treating the independent variables as random, while they are often designed variables (e.g., it seems unnatural to consider route or airline as random), and relatively poor practical performance in estimating the conditional distribution.

We instead focus on direct modeling of the unknown conditional distribution of transport risk y given predictors $\mathbf{x} = (x_1, \cdots, x_p)' \in \mathcal{X}$ ($\mathcal{X}$ is the sample space for the predictors $\mathbf{x}$) without specifying a model for the marginal of $\mathbf{x}$. In our context, predictors $\mathbf{x}$ = {airline (a), route (r), month (m), number of legs (leg), initial deviation (dev_{start}), planned duration (dur), cargo weight (wgt), cargo number of pieces (pcs)} (as specified in Table 1). In particular, we assume the transport risk y arises from a convolution

$$y \mid \mathbf{x} \sim \int k(y \mid \boldsymbol{\psi}) \, G_{\mathbf{x}}(\mathrm{d}\boldsymbol{\psi}) \tag{1}$$

where $k(\cdot \mid \boldsymbol{\psi})$ is a given parametric kernel indexed by parameters $\boldsymbol{\psi}$ (e.g., Normal kernel $k(\cdot \mid \boldsymbol{\psi})$ is indexed by $\boldsymbol{\psi}$ =(mean, standard deviation)), and the mixing distribution $G_{\mathbf{x}}$ is allowed to vary flexibly with predictors $\mathbf{x} \in \mathcal{X}$. The typical form in the BNP literature (refer to Rodriguez and Dunson (2011) for references) lets

$$G_{\mathbf{x}} = \sum_{l=1}^{L} \omega_l(\mathbf{x}) \, \delta_{\boldsymbol{\psi}_l(\mathbf{x})}, \text{ where } \sum_{l=1}^{L} \omega_l(\mathbf{x}) = 1 \text{ and } \omega_l(x) \geq 0 \tag{2}$$

where the atoms $\{\boldsymbol{\psi}_l(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}_{l=1}^{L}$ are $i.i.d$ sample paths from a stochastic process over $\mathcal{X}$, and $\{\omega_l(\mathbf{x}), \mathbf{x} \in \mathcal{X}\}$ are predictor-dependent probability weights that sum to one for all x. The above form is too general to be useful and it is necessary to make some simplifications for practical implementation. One common possibility is to introduce predictor dependence only in the $G_{\mathbf{x}}$ atoms, $\psi_l(x)$, while keeping weights, $\omega_l(\mathbf{x}) = \omega_l$, fixed. However, this approach tends to have relatively poor performance in our experience, including the air cargo transport risk data, compared with models that instead fix the atoms, while allowing the weights to vary.

In our case, the peak locations of the dependent variable, transport risk, are almost constant (i.e., daily peaks for international shipments, and some additional few-hourly peaks for domestic

shipments besides the daily peaks). However, the heights of the peaks change greatly along with $\mathbf{x}$ (e.g., route, airline, demand variables). The height of each peak represents (roughly) the probability for the observation to fall into the kernel centered around that peak. For example, if conditional on certain $\mathbf{x}_1$, the peak around 24 hours is relatively high, then a shipment, conditional on $\mathbf{x}_1$, has a large probability of being delayed for one day. On the other hand, if conditional on certain $\mathbf{x}_2$, there is only one peak around 0 high and visible, then a shipment, conditional on $\mathbf{x}_2$, probably arrives close to the planned arrival time. So, in our context, to find out how the height of each peak depends on $\mathbf{x}$ is of central interest.

Inducing dependence structure in the weights can be difficult and lead to complex and inefficient computational algorithms, limiting the applicability of the models. To overcome these difficulties, we adopt the PSBP mixture model, which has the advantages of computational tractability and consistency under weak regularity conditions.

3.2. Bayesian Probit Stick-breaking Process

Recalling the general form of the mixing measure in Equation (2), ***stick-breaking*** weights are defined as $\omega_l = u_l \prod_{p<l}(1-u_p)$, where the stick-breaking ratios are independently distributed $u_l \sim H_l$ for $l < L$ and $u_L = 1$ for the case of finite L. In the baseline case in which there is no predictor, ***Probit stick-breaking*** weights are constructed as

$$u_l = \Phi(\gamma_l),\ \gamma_l \sim \mathsf{N}(\mu, \phi)$$

where $\Phi(\cdot)$ denotes the cumulative distribution function (cdf) for the standard normal distribution. μ is the mean and ϕ is the precision (the reciprocal of the variance) of a normal distribution such that for $x \sim \mathsf{N}(\mu, \phi)$, the probability density function (pdf) is $f(x) = \sqrt{\frac{\phi}{2\pi}} \exp\left\{-\frac{\phi}{2}(x-\mu)^2\right\}$. For a finite L, the construction of the weights ensures that $\sum_{l=1}^{L} \omega_l = 1$. When $L = \infty$, $\sum_{l=1}^{\infty} \omega_l = 1$ almost surely (Rodriguez and Dunson 2011).

The use of Probit transformation to define the weights builds a mapping between a real number γ_l from $-\infty$ to $+\infty$ into $u_l \in (0,1)$. Thus, the transformation allows researchers to restate the model using normally distributed latent variables γ_l, facilitating computation via data augmentation Gibbs

sampling algorithms presented in §3.4. This transformation also makes model extensions to include additional structure (e.g,. predictors) straightforward. Additionally, the Probit transformation simplifies prior elicitation as presented at the end of §3.2.

In order to make $\omega_l(\mathbf{x})$ predictor-dependent, we further express the latent variables γ_l as a linear regression function of $\mathbf{x}$, $\{\gamma_l(\mathbf{x}), \mathbf{x} \in \mathcal{X}\}$ (In this paper we use superscript as an index rather than the exponent of the parameter):

$$\omega_l(\mathbf{x}) = \Phi(\gamma_l(\mathbf{x})) \prod_{p<l} (1 - \Phi(\gamma_p(\mathbf{x}))) \tag{3}$$

$$\begin{aligned}\gamma_l(\mathbf{x}) = {} & \theta_l^1 + \theta_a^2 + \theta_r^3 + \theta_{(a,r)}^4 + \theta_m^5 + \theta_{leg}^6 + \theta_{(a,leg)}^7 + f_1\left(dev_{start} \mid \boldsymbol{\theta}^8\right) \\ & + f_2\left(dur \mid \boldsymbol{\theta}^9\right) + f_3\left(\log(wgt) \mid \boldsymbol{\theta}^{10}\right) + f_4\left(\log(pcs) \mid \boldsymbol{\theta}^{11}\right)\end{aligned} \tag{4}$$

where $\{\theta_l^1\}$ controls the baseline probability of latent class l ($l = 1, \cdots, L$), $\{\theta_a^2\}$ controls the baseline heterogeneity of airline a ($a = 1, \cdots, 20$), $\{\theta_r^3\}$ controls the heterogeneity of route r ($r = 1, \cdots 1336$), $\left\{\theta_{(a,r)}^4\right\}$ represents the dependence of weights on possible interactions between airlines and routes, and the meanings of $\{\theta_m^5\}$, $\{\theta_{leg}^6\}$,$\left\{\theta_{(a,leg)}^7\right\}$ are similar. In addition, f_1, f_2, f_3 and f_4 are spline functions expressed as a linear combination of B-splines of degree 4, where the knots of dev_{start} are [-3, -2, -1, 0, 1, 2, 3], the knots of dur are [1, 2, 4, 6, 8, 10], the knots of $log(weight)$ are [2, 4, 6, 8] and the knots of $log(pcs)$ are [1, 3, 5]. Here we use the logarithm form of cargo weight (wgt) and number of pieces (pcs) as the predictors, since the original distributions are highly skewed. To ensure identification of the parameters, we let $\theta_1^2 = \theta_1^3 = \theta_{(1,r)}^4 = \theta_{(a,1)}^4 = \theta_1^5 = \theta_{(1,leg)}^6 = \theta_{(a,1)}^7 = 0$ for all a, r and interactions in sample space $\mathcal{X}$.

3.2.1. Prior for Parameters in Weight To retain conjugacy, we choose Normal priors for parameters $\Theta = \{\{\theta_l^1\}, \{\theta_a^2\}, \{\theta_r^3\}, \left\{\theta_{(a,r)}^4\right\}, \{\theta_m^5\}, \left\{\theta_{leg}^6\right\}, \left\{\theta_{(a,leg)}^7\right\}, \boldsymbol{\theta}^8, \boldsymbol{\theta}^9, \boldsymbol{\theta}^{10}, \boldsymbol{\theta}^{11}\}$

$$\theta_j^i \sim \mathsf{N}\left(\nu^i, \epsilon^i\right), \text{ for } i = 8, \cdots, 11 \text{ and } j = 1, \cdots, n(i)$$

where $n(i)$ is the number of B-spline basis used for predictor i. For the coefficients of 7 categorical independent variables $\boldsymbol{\theta}^1, \cdots, \boldsymbol{\theta}^7$ (i.e., $\boldsymbol{\theta}^1 = \{\theta_l^1\}$ etc), we build a hierarchy, which enables information borrowing among parameters in one category

$$\theta_l^1 \sim \mathsf{N}\left(\Phi^{-1}\left(\frac{1}{L-l+1}\right), \epsilon^1\right), \; \theta_a^2 \sim \mathsf{N}(0, \epsilon^2), \cdots \; \theta_{(a,leg)}^7 \sim \mathsf{N}\left(0, \epsilon^7\right).$$

where $\epsilon^i \sim \mathsf{G}(c_i, d_i)$ for $i = 1, 2, \cdots, 7$. Here $\mathsf{G}(a, b)$ is a Gamma distribution such that for $x \sim \mathsf{G}(a, b)$ the pdf is $f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$. We use the specially designed prior of θ_l^1 to enforce the same prior baseline probability of each cluster $l = 1, 2, \cdots, L$.

3.3. Normal Kernel

A mixture of a moderate number of Normals is known to produce an accurate approximation of any smooth density. Also motivated by computational tractability of the Normal distribution (e.g., through conjugacy in posterior calculations), we specify the parametric kernel, $k(\cdot \mid \boldsymbol{\psi})$, of PSBP mixture model as Normal distribution, $\mathsf{N}(\mu, \phi)$, where $\boldsymbol{\psi} = (\mu, \phi)$. Recalling that our mixture model takes the form in Equation (1), we replace the kernel in the above equation with Normal and use the PSBP specified prior $G_{\mathbf{x}}$. Then the conditional distribution of y can be expressed in the simple form

$$y \mid \mathbf{x} = \sum_{l=1}^{L} \omega_l(\mathbf{x}) \, \mathsf{N}(y \mid \mu_l, \phi_l)$$

3.3.1. Prior for Parameters in Normal Kernel The prior of atoms $\{(\mu_l, \phi_l), l = 1, 2, \cdots, L\}$ is $\mathsf{NG}(\zeta_\mu, \xi_\mu, a_\phi, b_\phi)$, a conjugate Normal-Gamma prior such that

$$\mu_l \sim \mathsf{N}(\zeta_\mu, \xi_\mu \phi_l), \quad \phi_l \sim \mathsf{G}(a_\phi, b_\phi).$$

where $l = 1, 2, \cdots, L$.

3.4. Posterior Computation

Bayesian posterior sampling can be challenging for a large model like ours, which involves more than 2000 parameters. General Markov chain Monte Carlo (MCMC) schemes, such as Metropolis-Hastings, have limited capability in a high-dimensional setting. This is because they require selection of a large number of algorithmic tuning parameters in proposal distributions, and this selection needs to be carefully done to obtain adequate computational efficiency. For this reason, we adopt blocked Gibbs sampling algorithms. These algorithms simultaneously update vectors of parameters by sampling from their conditional distribution, which avoid algorithm tuning and typically have superior mixing. In order to obtain tractable conditional distributions for sampling, a common strategy is data augmentation.

Specifically, we augment the observed data with latent variables, thus obtain a simple Gibbs sampling algorithm for posterior computation of model parameters (Rodriguez et al. 2009). First we focus on case when $L < \infty$. For each observation $y_j \mid \mathbf{x}$, (corresponding to replicate j conditional on $\mathbf{x}$, $j = 1, \cdots, n(\mathbf{x})$ if there are $n(\mathbf{x})$ replicates, otherwise j is dropped if there are no replicates, i.e., $n(\mathbf{x}) = 1$), we introduce a latent indicator variable $s_j(\mathbf{x})$ such that $s_j(\mathbf{x}) = l$ if and only if observation $y_j \mid \mathbf{x}$ is sampled from mixture component l $(l = 1, 2, \cdots, L)$. The use of these latent variables is standard in mixture models.

With the help of latent indicators $s_j(\mathbf{x})$, Gibbs sampling of more than 2000 model parameters can be classified into four categories, as presented in the following four subsections.

3.4.1. Gibbs Sampling for Kernel Parameters We use "$\cdots$" to indicate *all the other parameters and data*. The full conditional distribution of the component-specific parameters, μ_l and ϕ_l, is given by

$$p(\mu_l, \phi_l \mid \cdots) \propto \mathsf{NG}(\mu_l, \phi_l \mid \zeta_\mu, \xi_\mu, a_\phi, b_\phi) \prod_{(\mathbf{x},j)\ s.t.\ s_j(\mathbf{x})=l} \mathsf{N}(y_j \mid \mu_l, \phi_l)$$

where $\propto$ represents "proportional to", NG is the Normal-Gamma conjugate prior of μ_l and ϕ_l. Simplified by the conjugacy structure, the Gibbs sampling of kernel mean μ_l is carried out by

$$\mu_l \mid \cdots \sim \mathsf{N}\left(\left[\zeta_\mu + n_l\phi_l\right]^{-1}\left[\zeta_\mu\xi_\mu + h_l\phi_l\right], \xi_\mu + n_l\phi_l\right)$$

where $n_l = \sum_{\mathbf{x}\in\mathcal{X}} \sum_{j=1}^{n(\mathbf{x})} \mathbf{1}_{(s_j(\mathbf{x})=l)}$ and $h_l = \sum_{\mathbf{x}\in\mathcal{X}} \sum_{j=1}^{n(\mathbf{x})} y_j(\mathbf{x}) \mathbf{1}_{(s_j(\mathbf{x})=l)}$. Similarly, the Gibbs sampling of kernel precisions ϕ_l is

$$\phi_l \mid \cdots \sim \mathsf{G}\left(a_\phi + \frac{n_l}{2}; b_\phi + \frac{1}{2}\sum_{\mathbf{x}\in\mathcal{X}}\sum_{j=1}^{n(\mathbf{x})} \left(y_j(\mathbf{x}) - \mu_l\right)^2 \mathbf{1}_{(s_j(\mathbf{x})=l)}\right)$$

3.4.2. Gibbs Sampling for Weight Parameters: Latent Indicators Conditional on kernel parameters and the realized values of the weights $\{\omega_l(\mathbf{x}), \forall \mathbf{x} \in \mathcal{X}\}_{l=1}^{L}$, the distribution of the indicators is multinomial with probability given by

$$\Pr(s_j(\mathbf{x}) = l \mid \cdots) \propto \omega_l(\mathbf{x}) \mathsf{N}(y_j(\mathbf{x}) \mid \mu_l, \phi_l),$$

So we can sample $s_j(\mathbf{x})$ $(j = 1, \cdots, n(\mathbf{x}))$ from a multinomial conditional distribution:

$$Pr(s_j(\mathbf{x}) = l \mid \cdots) = \frac{\omega_l(\mathbf{x}) \mathsf{N}(y_j(\mathbf{x}) \mid \mu_l, \phi_l)}{\sum_{p=1}^{L} \omega_p(\mathbf{x}) \mathsf{N}(y_j(\mathbf{x}) \mid \mu_l, \phi_l)}$$

3.4.3. Gibbs Sampling for Weight Parameters: Latent Auxiliary Variable In order to sample the latent processes $\{\gamma_l(\mathbf{x}), \forall \mathbf{x} \in \mathcal{X}\}_{l=1}^{L}$ and the corresponding weights $\{\omega_l(\mathbf{x}), \forall \mathbf{x} \in \mathcal{X}\}_{l=1}^{L}$, we ***augment*** the data with a collection of conditionally independent latent variables $z_{jl}(\mathbf{x}) \sim \mathsf{N}(\gamma_l(\mathbf{x}), 1)$ $(j = 1, \cdots, n(\mathbf{x}))$. We aim to make the probability of observing $\{z_{j1}(\mathbf{x}), \cdots, z_{jl}(\mathbf{x})\}$ equal to $\omega_l(\mathbf{x})$ in Equation (3), thus the event of observing $\{z_{j1}(\mathbf{x}), \cdots, z_{jl}(\mathbf{x})\}$ can represent the event of observing $s_j(\mathbf{x}) = l$ as denoted at the beginning of §3.4. Specifically if $z_{jp}(\mathbf{x}) < 0$ for all $p < l$ and $z_{jl}(\mathbf{x}) > 0$, we define $s_j(\mathbf{x}) = l$. For a finite L case, we define $s_i(\mathbf{x}) = L$ if $z_{ip}(\mathbf{x}) < 0$ for all $p \leq L - 1$. Then we have

$$\begin{aligned}\Pr(s_j(\mathbf{x}) = l) &= \Pr(z_{jl}(\mathbf{x}) > 0, z_{jp}(\mathbf{x}) < 0 \text{ for } p < l) \\ &= \Phi(\gamma_l(\mathbf{x})) \prod_{p<l} \{1 - \Phi(\gamma_p(\mathbf{x}))\}\end{aligned}$$

independently for $j = 1, \cdots, n(\mathbf{x})$. In this way, $\Pr(s_j(\mathbf{x}) = l)$ equals to $\omega_l(\mathbf{x})$ as defined in Equation (3). This data augmentation scheme simplifies computation as it allows us to implement the following Gibbs sampling scheme

$$z_{jl}(\mathbf{x}) \mid \cdots \sim \mathsf{N}(\gamma_l(\mathbf{x}), 1)\, \mathbf{1}_{\mathbf{\Omega}_l},\ \forall l \leq \min\{s_j(\mathbf{x}), L - 1\},$$

with

$$\mathbf{\Omega}_l = \begin{cases} \{z_{jl}(\mathbf{x}) < 0\}, & \text{if } l < s_j(\mathbf{x}), \\ \{z_{jl}(\mathbf{x}) \geq 0\}, & \text{if } l = s_j(\mathbf{x}) < L \end{cases}$$

where $\mathsf{N}(\cdot)\, \mathbf{1}_{\Omega}$ denotes a normal distribution truncated to the set Ω.

3.4.4. Gibbs Sampling for Weight Parameters: Latent Processes The latent process $\{\gamma_l(\mathbf{x}), \forall \mathbf{x} \in \mathcal{X}\}$ is built on parameters $\Theta = \{\{\theta_l^1\}, \{\theta_a^2\}, \{\theta_r^3\}, \left\{\theta_{(a,r)}^4\right\}, \{\theta_m^5\}, \{\theta_{leg}^6\}, \left\{\theta_{(a,leg)}^7\right\}, \boldsymbol{\theta}^8, \boldsymbol{\theta}^9, \boldsymbol{\theta}^{10}, \boldsymbol{\theta}^{11}, \boldsymbol{\theta}^{12}\}$ and hyper-parameters $\Upsilon = \{\epsilon^i, \forall i = =1, 2, \cdots, 7\}$. The distribution of Θ and Υ, conditional on the augmented data, is given by

$$p(\Theta, \Upsilon \mid \cdots) \propto \left[\prod_{\mathbf{x},j} p\left(\mathbf{z}_j(\mathbf{x}) \mid \boldsymbol{\gamma}_j(\mathbf{x})\right)\right] p(\Theta) p(\Upsilon)$$

where $p(\Theta)$ is the prior distribution of Θ and $p(\Upsilon)$ is the prior distribution of Υ, and $j = 1, \cdots, n(\mathbf{x})$. The posterior sampling can be easily implemented by taking advantage of the normal priors we

choose. Due to similarities of the Gibbs sampling schemes for Θ and Υ, here we only give updating schemes for two examples: one for coefficients $\{\theta_l^1\}_{l=1}^L \in \Theta$ and the other one for hyper-parameter $\epsilon^1 \in \Upsilon$.

1. For θ_l^1 $(l = 1, 2, \cdots, L)$, the posterior Gibbs sampling follows normal distribution given by

$$\theta_l^1 \mid \cdots \propto \mathsf{N}\left(\mu_{\theta_l^1}, \phi_{\theta_l^1}\right)$$

where $\mu_{\theta_l^1} = \left\{\Phi^{-1}\left(\frac{1}{L-l+1}\right) + \sum_{\mathbf{x}\in\mathcal{X}}\sum_j \left[z_{jl}(\mathbf{x}) - \Delta_{jl}(\mathbf{x})\right]\mathbf{1}\left(s_j(\mathbf{x}) \geq l\right)\right\} / (n_l + 1)$, $\phi_{\theta_l^1} = (n_l + \epsilon^1) / (n_l + 1)$, $n_l = \sum_{\mathbf{x}\in\mathcal{X}}\sum_j \mathbf{1}\left(s_j(\mathbf{x}) \geq l\right)$ and $\Delta_{jl}(\mathbf{x}) = (\gamma_j(\mathbf{x}) - \theta_l^1)\,\mathbf{1}\left(s_j(\mathbf{x}) \geq l\right)$.

2. For ϵ^1 the posterior Gibbs sampling follows Gamma distribution given by

$$\epsilon^1 \mid \cdots \propto \mathsf{G}\left(c_1 + \frac{L}{2}, d_1 + \frac{\sum_{l=1}^L \theta_l^1 \cdot \theta_l^1}{2}\right)$$

In the case $L = \infty$, we can easily extend this algorithm to generate a slice sampler, as discussed in Papaspiliopoulos (2008). Alternatively, the results in Rodriguez and Dunson (2011) suggest that a finite PSBP with a large number of components (30 – 40, depending on the value of μ) can be used instead (Ishwaran and Zarepour 2002). So we use $L = 50$ as the number of components in this paper, the provides a conservative upper bound as many of these components may not be utilized.

3.5. Model Fitting Assessment

We use three methods to assess model fitting: cross validation, posterior predictive checking and visual inspection. Cross validation is widely used to check the out-of-sample predictive capability of a model and also limits problems like overfitting. Specifically, we use a 3-fold cross validation based on predictive log likelihood, a strictly proper scoring rule for density forecasts (Gneiting and Raftery 2007). The model with highest predictive log-likelihood is shown in Equation (5)

$$\gamma_l(\mathbf{x}) = \theta_l^1 + \theta_a^2 + \theta_r^3 + \theta_{(a,r)}^4 + \theta_m^5 + \theta_{leg}^6 + f_1\left(dev_{start} \mid \boldsymbol{\theta}^8\right) + f_2\left(dur \mid \boldsymbol{\theta}^9\right) + f_3\left(\log(wgt) \mid \boldsymbol{\theta}^{10}\right) \quad (5)$$

where predictors $log(pcs)$ and airline-leg interactions are dropped. All the following analyses are based on estimation of the model specified by Equation (5).

We compare this model with OLS, which is widely used in the previous research of flight delays. First, we replicate two data sets predicted by OLS and our model separately, and compare them with the original data. The basic idea of posterior predictive checking is that if the model specification is appropriate, we would expect to see something similar to the real data (Rubin 1984). The replicated data by OLS resembles the shape of real data poorly. Whereas, the replicated data by our model resembles the real data very well. Next, we define several test statistics to test our model compared against real data and the OLS prediction. Specifically, we define test statistics: (1) mean, (2) standard deviation, (3) $\mathsf{Prob}\,(y < -24 \text{ or } y \geq 36 \mid \Psi)$, (4) $\mathsf{Prob}\,(-24 \leq y < 36 \mid \Psi)$ (see Figure **??**). Obviously, OLS largely underestimates extreme situations like more than 24 hours earliness or more than 36 hours delays, while overestimating recurrent risk such as deviations between -24 to 36 hours. OLS concentrates at mean estimation, whose prediction is almost the same with data mean, while the standard deviation is substantially underestimated. The posterior predictive statistics by PSBP are close to the true values.

We further check model fitting at a more granular level – airline-route level. In Figure 3, the histogram is drawn from real data, the solid line is the predictive conditional density by PSBP (the posterior 95% probability intervals are too narrow to be visible in this figure), while the dashed line is predicted by OLS. PSBP captures the location and weights of peaks accurately while OLS predicts badly.

4. Results

Table A.2 in Appendix §A.2 shows the posterior mean and 95% probability interval of (selected) model parameters. There are several things to note from the table:

1. The 50 kernel means, $\mu_1, \mu_2, \cdots, \mu_{50}$, range from -70.0 to 77.5 (hours), indicating the model predicted deviation concentrates within -3 to 3 days, consistent with the data. The 50 kernel standard deviations, $1/\sqrt{\phi_1}, 1/\sqrt{\phi_2}, \cdots, 1/\sqrt{\phi_{50}}$, range from 0.62 to 84.4, meaning the Normal kernels can be very narrow or flat, allowing for flexible estimation.

2. Level parameters, $\theta_1^1, \theta_2^1, \cdots, \theta_{49}^2$, vary from -10.9 to 6.74, and the wide range suggests strong variation in risk. For example, if an airline-route pair has $\gamma_l(\mathbf{x}) - \theta_l^1$ close to zero, then for certain

l with θ_l^1 smaller than -5, the weight $\propto \Phi(\gamma_l(\mathbf{x})) \approx \Phi(-5) \approx 0$, thus eliminating the inclusion of this component. By similar arguments, θ_l^1 can also help determine, for which $\gamma_l(\mathbf{x}) - \theta_l^1$, component l plays major role.

3. The posterior distributions of coefficients all present substantial learning from their prior distribution; in addition, the 95% probability intervals are narrow.

4. The posterior estimation of airline coefficient (we disguise the names of airlines for confidential reasons. The airline index used here is randomly assigned), θ_a^2, shows great heterogeneity, and the large standard deviation, $1/\sqrt{\epsilon^2}$, which measures the variations among airlines, confirms this from one other aspect. Closer inspection reveals that except A1, whose coefficient is fixed at zero for identification, 18 of the remaining 19 airlines' 95% probability intervals don't include 0. Furthermore, many of them are far from zero, implying large impact on transport risk. However, based on OLS, only 2 of the 19 airlines are significantly different from 0 at 5% confidence level. This huge difference underlies the principle of the two estimation methods. OLS focuses in estimating the effects of independent variables on distribution ***mean***, and its results indicate airlines don't necessarily affect the mean of transport risk much. However, PSBP's results show that airlines are playing an important role on selecting and weighting possible kernels, which affects the tail shape, number of peaks, probability of extreme observation etc. These results and comparison once again show that OLS, which cannot detect the airlines' (and some other predictors' including routes' etc) impact on transport risk in this case, would lose considerable valuable information.

5. Since the number of routes and their interactions with airline are large, 1336 and 587 respectively, we don't include their posterior summaries in Table A.2. However, posterior summaries of hyper-parameters standard deviation, $1/\sqrt{\epsilon^3}$, illustrate the large heterogeneity between routes. More importantly, the large standard deviation, $1/\sqrt{\epsilon^4}$, represents possibly huge differences in terms of the distribution of transport risks on the same route while by different airlines. This suggests that a careful selection of carriers can result in dramatically different shipping experiences.

5. Applications

Estimates of predictive conditional probability density functions (Cpdf) is key to generating data-driven operations strategies. In this section, we provide several examples for how posterior Cpdf can aid decision making. We note that there are other applications of our transport risk models.

5.1. Service Comparison for One Shipment

The most straightforward use of PSBP posterior estimation is to provide predictive Cpdf of transport risk to shippers based on their predetermined demand variables and selectable decision variables (see Table 1). This not only helps the shipper to find a preferable service but also helps the forwarder to set a price quote. Assume a customer comes with predetermined demand requirement $c = \{r, m, wgt\}$ and is choosing from services $s = (a, leg, dur) \in S(c)$, where $S(c)$ is the set of services available given c. Here, even though the initial deviation, dev_{start}, is one of the decision variables, we set it to 0 because this variable is unknown and not selectable before shipping starts. Let $f(risk \mid c, s)$ be the predictive distribution of transport risk conditional on c and a chosen s, and $l_i(risk)$ be customer i's loss function. The optimal conditional choice of s, which minimizes expected transport loss, is defined as

$$\begin{aligned} (s \mid c)_i^* &\triangleq \operatorname{argmin}_{s \in S(c)} Loss_i(s \mid c) \\ Loss_i(s \mid c) &= \int l_i(risk)\, f(risk \mid c, s)\, \mathrm{d}dev \end{aligned} \tag{6}$$

where $Loss_i(s \mid c)$ is customer i's expected loss of choosing s given c. Estimating each customer's unknown loss function $l_i(dev)$ is another interesting study of practical value, but is outside the scope of this paper. Here we use several generic loss functions to illustrate how to use predicted $f(risk \mid c, s)$ to aid service selection.

In Figure 4 are 6 choices as shown by the figure titles, on the route from Frankfurt to Atlanta. The choices are randomly picked from the data. We use the following three loss functions:

$$l_1(risk) = C_1 \cdot risk \qquad l_2(risk) = C_2 \cdot \mathbf{1}\{risk > 18\} \qquad l_3(risk) = C_3 \cdot risk^2$$

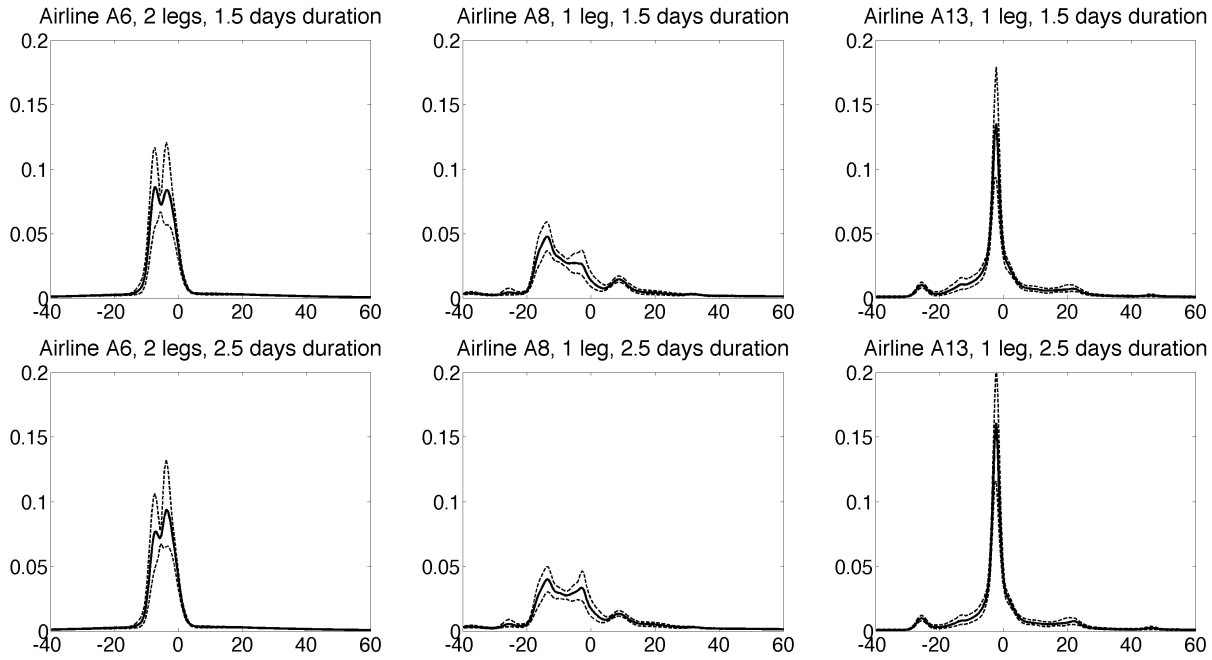

Figure 4 **From Frankfurt (Germany) to Atlanta (United States)**

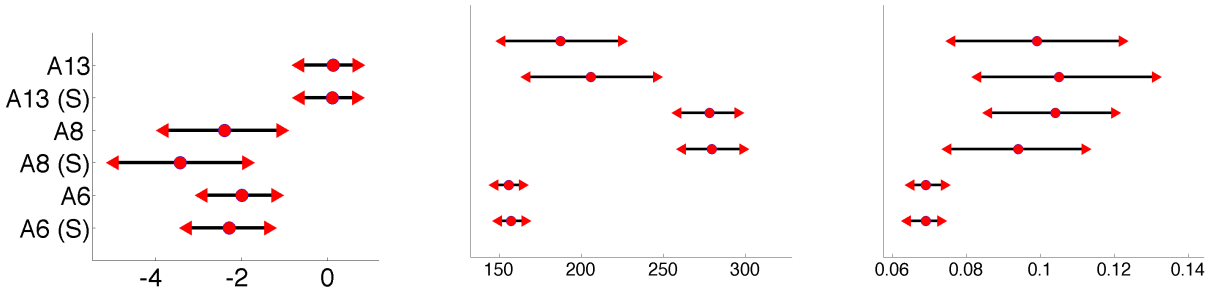

Figure 5 **Ranking based on expected transport risk**

l_1 naturally arises when a risk neutral shipper is adverse to delays while fond of early arrivals; l_2 is more proper when a shipper is sensitive to extreme delays exceeding certain threshold (18 hours in our example); l_3 is used when a shipper is risk adverse and dislikes any deviations from the plan, neither negative nor positive. Under these loss functions, the expected losses have simple analytical forms

$$Loss_1 = C_1 \cdot \mathsf{E}_f \qquad Loss_2 = C_2 \cdot (1 - F(18)) \qquad Loss_3 = C_3 \cdot \left(\mathsf{Var}_f + \mathsf{E}_f^2\right)$$

where f is short for $f(risk \mid c, s)$ and F is the corresponding cumulative density function. Figure 5

presents the expected losses (with posterior 95% probability intervals) calculated for the six choices under 3 risk functions with $C_1 = C_2 = C_3 = 1$, in which we use (S) to indicate *speedy* service. We observe (1) the rank of services in terms of expected loss varies by loss functions; (2) choice of airlines is playing a more dominant role than the choice between normal and speedy services given an airline.

With estimated expected loss of each choice, forwarders can offer different price quotes to different types of shippers. In this example, a forwarder can increase revenue by lowering A8's prices to attract price-sensitive shippers and increasing A6's prices to attract quality-sensitive shippers under loss function 2.

5.2. Supplier Ranking on Route or Higher Level

Unlike a shipper, whose decision is made at the level of each shipment, a forwarder plans its business at the route or higher level. To help solve problems at high levels, the full predictive Cpdf should be integrated. Specifically, let the full information set be $U = \{a,\ r,\ m,\ leg,\ dur,\ dev_{start},\ wgt\}$, for $U = U_1 \cup U_2$ and $U_1 \cap U_2 = \phi$, then

$$f\left(risk \mid U_1\right) = \int f\left(y \mid U_1, U_2\right) f\left(U_2\right) dU_2$$

where U_1 contains variables of central interest, and other variables in U_2 are integrated out. For example, a practical problem faced by a forwarder is whether to choose a carrier on a certain route and how much capacity to reserve from it. For such decisions, an estimation of the airline's service reliability is a critical input. In this case airlines and routes are of interest, so we let $U_1 = \{a, r\}$ and $U_2 = U - U_1$. By using Equation (6) with c and s replaced by r and a, the forwarder can obtain expected losses by each airline $a \in S\left(r\right)$, which, in turn, can help make the right capacity reservation and pricing decisions.

5.3. Baseline Comparison

Our result can also be used to generate baseline comparisons of various factors. Baseline effect of a certain factor excludes the effects of any other factors, thus allowing for a direct comparison between factors of one type. One interesting example is to understand the baseline performance of each

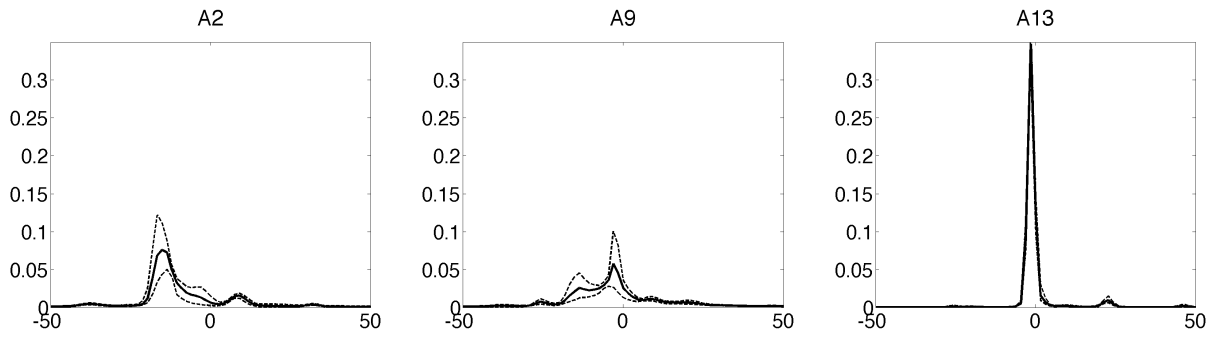

Figure 6 Sample airline reference performances

airline, in which case a direct comparison is impossible due to the fact that airlines serve different routes. To achieve this baseline comparison, we use the average value for all other predictors, except airline effects θ_a^2, as their reference levels. Then we plug these reference levels in the posterior samples of each airline and then obtain the reference risk distribution for each airline (See Figure 6 for 3 samples from the 20 airlines. Due to space limitation, the remaining 17 baseline distributions are not included in this paper). From the plots we can directly compare airlines, which differ from each other by the number, locations, and heights of peaks. As such, our model allows baseline comparison based on distribution knowledge. This offers a much richer comparison than those appearing in the literature based on single average metrics. Meanwhile, the richer tool allows us to obtain simple metric comparisons as special cases.

For example, using a U.S. passenger flight data set, Deshpande and Arikan (2012) analyzed single-leg flight truncated block time, which is transport risk plus planned duration minus initial deviation. Initial deviation is defined as the positive delay of the previous flight by the same craft if applicable and zero otherwise. The authors argue that if the truncated block time is shorter than the scheduled block time, the airline incurs an overage cost of C_o per unit overage time. Otherwise, the airline incurs an underage cost C_u per unit shortage time. The authors then estimate the overage to underage ratio, $\varphi = C_0/C_u$, for each flight, and calculate the mean ratio of flights served by a certain airline as the airline-wise overage to underage ratio, φ_a. Using our international air cargo data, we can obtain an analogous metric by replacing "schedule block time" and "truncated block time" in their paper with dur and ($dur +$ arrival deviation $- [dev_{start}]^+$). One concern of estimating

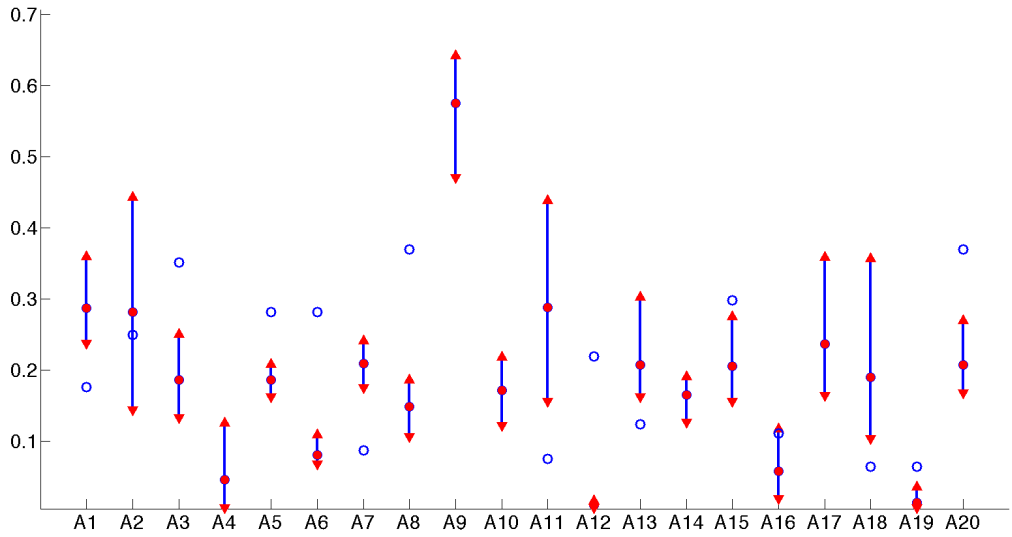

Figure 7 **Overage to underage ratio of airlines**

airline-wise ratio φ_a by simply calculating the average of flight-wise ratios is that the effects from other factors, such as routes etc, cannot be excluded. Thus, the calculated overage to underage ratio of each airline, φ_a, cannot be used for direct comparison of airlines' intrinsic service quality. Baseline distribution of airlines, on the other hand, is a good solution to this problem. Specifically, the optimal dur^* is defined by news-vendor solution that

$$\begin{aligned} \mathsf{Prob}\left(dur^* + \text{arrival deviation} - [dev_{start}]^+ \le dur^* \mid a\right) &= \frac{1}{1+\varphi_a} \\ \mathsf{Prob}\left(\text{arrival deviation} \le 0 \mid a\right) &= \frac{1}{1+\varphi_a} \end{aligned} \tag{7}$$

where we use the fact that the reference level of $[dev_{start}]^+$, calculated by the data average, is zero. Thus each airline's overage to underage ratio is calculated by $\varphi_a = \frac{1}{F_a(0)} - 1$; see Figure 7 for the calculated overage to underage ratios of 20 airlines with 95% probability intervals.

The overage to underage ratio φ_a is related to airline's on-time probability by Equation 7: the higher the on-time rate the lower the ratio φ_a. We compare our results to C2K Monthly Statement issued by IATA. In particular, we choose monthly report issued in November 2012, the same period of our data, and convert the reported airlines' on-time rates into their overage/underage ratios (represented by the circles in Figure 7). The circles deviate from our estimations, the solid dots,

following no obvious rules. We believe this is because IATA calculated the on-time rate by simply averaging on-time times of an airline, which fails to exclude the impacts from factors other than the airline, e.g., cargo weight, route, and thus results in unfair comparison. The baseline distribution we calculated can also be used to calculate many other metrics, such as variance, probability of extreme disruptions etc, rather than the simple on-time rate reported by IATA's monthly report.

6. Conclusions and Future Directions

Using data from international air cargo logistics, we investigate ways to assess and forecast transport risks, defined as the deviation between actual arrival time and planned arrival time. To accommodate the special multimodal feature of the data, we introduce a Bayesian nonparametric mixture model, the Probit stick-breaking process (PSBP) mixture model, for flexible estimation of conditional density function of transport risk. Specifically, we build a linear structure, including demand variables and decision variables, into kernel weights so that the probability weights change with predictors. Advantages of the PSBP include its generality, flexibility, relatively simple sampling algorithm and consistency under weak regulation conditions. Our results show that this method achieves accurate forecasts, while the simpler OLS method can lead to misleading inferences. We also demonstrate how an accurate estimation of transport risk Cpdf can help shippers to choose from multiple available services, and help a forwarder to set targeting price, etc. In addition, we show how to use the model to estimate baseline performance of a predictor, such as an airline. We compare our findings with performance reports issued by IATA and point out the shortcomings of IATA's simple way of ranking airlines. We note that the usage of our method can be much broader than the examples shown here. Indeed, any decisions involving a distribution function needs an estimated Cpdf.

Our study serves as a stepping stone to deeper studies in the air cargo transport industry, or more generally, the transportation industry, which generates tons of data everyday yet lacks proper techniques for data analysis. According to a 2011 McKinsey report (Manyika et al. 2011), in the transportation and warehousing sector, the main focus of our paper, IT intensity is among the top 20% and data availability is among the top 40% of all sectors, but the data-driven mind-set is merely

at the bottom 20%. The authors' communication with leaders in this industry, from whom we get the data supporting this research project, confirms this situation, "... we have plenty of data, or we could say we have all the data possible, but we don't know how to use the data...".

One of the interesting findings of our paper is that airlines have critical impact on the shape of the transport risk distribution rather than the mean focused on by OLS. For future research, we hope to be able to obtain information regarding why and how airlines are performing so differently on the same routes. By knowing the root drivers of airline service performance, the service quality can be improved for each airline rather than simply choose the best performer.

References

Chung, Y, DB Dunson. 2009. Nonparametric Bayes conditional distribution modeling with variable selection. *J. Amer. Statist. Assoc.* **104**(488) 1646–1660.

Deshpande, V, M Arikan. 2012. The impact of airline flight schedules on flight delays. *Manufacturing Service Oper. Management* **14**(3) 423–440.

Escobar, MD, M West. 1995. Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90**(430) 577–588.

Fan, J, Q Yao, H Tong. 1996. Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika* **83**(1) 189–206.

Federgruen, A, N Yang. 2009. Optimal supply diversification under general supply risks. *Oper. Res.* **57**(6) 1451–1468.

Gneiting, T, AE Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102**(477) 359–378.

Hall, P, RCL Wolff, Q Yao. 1999. Methods for estimating a conditional distribution function. *J. Amer. Statist. Assoc.* **94**(445) 154–163.

Hyndman, RJ, Q Yao. 2002. Nonparametric estimation and symmetry tests for conditional density functions. *J. Nonparametric Statistics* **14**(3) 259–278.

IATA. 2014a. C2k master operating plan. URL `http://www.iata.org/whatwedo/cargo/cargo2000/Pages/master-operating-plan.aspx`.

IATA. 2014b. Cargo 2000. URL http://www.iata.org/whatwedo/cargo/cargo2000/Pages/index.aspx.

Ishwaran, H, M Zarepour. 2002. Dirichlet prior sieves in finite normal mixtures. *Statistica Sinica* **12**(3) 941–963.

Kleindorfer, PR, JC Belke, MR Elliott, K Lee, RA Lowe, HI Feldman. 2003. Accident epidemiology and the US chemical industry: Accident history and worst-case data from RMP* info. *Risk Anal.* **23**(5) 865–881.

Kleindorfer, PR, GH Saad. 2005. Managing disruption risks in supply chains. *Production Oper. Management* **14**(1) 53–68.

Li, J, N Granados, S Netessine. 2014. Are consumers strategic? structural estimation from the air-travel industry. *Management Sci.* **60**(9) 2114–2137.

Lo, AY. 1984. On a class of Bayesian nonparametric estimates: I. density estimates. *Ann. Statistics* **12**(1) 351–357.

Manyika, J, M Chui, B Brown, J Bughin, R Dobbs, C Roxburgh, A Byers. 2011. Big data: the next frontier for innovation, competition, and productivity. Tech. rep., McKinsey Global Institute.

Morrell, PS. 2011. *Moving boxes by air: The economics of international air cargo*. Ashgate Publishing Company, Burlington, VT.

Mueller, ER, GB Chatterji. 2002. Analysis of aircraft arrival and departure delay characteristics. *AIAA aircraft technology, integration and operations (ATIO)*. Los Angeles, CA.

Muller, P, A Erkanli, M West. 1996. Bayesian curve fitting using multivariate normal mixtures. *Biometrika* **83**(1) 67–79.

Papaspiliopoulos, O. 2008. A note on posterior sampling from Dirichlet mixture models. Tech. rep., University of Warwick, Covertry, UK.

Pati, D, DB Dunson, ST Tokdar. 2013. Posterior consistency in conditional distribution estimation. *J. Multivariate Anal.* **116** 456–472.

Rodriguez, A, DB Dunson. 2011. Nonparametric Bayesian models through Probit stick-breaking processes. *Bayesian Anal.* **6**(1) 145–177.

Rodriguez, A, DB Dunson, J Taylor. 2009. Bayesian hierarchically weighted finite mixture models for samples of distributions. *Biostatistics* **10**(1) 155–171.

Rubin, DB. 1984. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statistics* **12**(4) 1151–1172.

Shumsky, RA. 1995. Dynamic statistical models for the prediction of aircraft take-off times. Ph.d. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Song, JS, PH Zipkin. 1996. Inventory control with information about supply conditions. *Management Sci.* **42**(10) 1409–1419.

Tomlin, B. 2006. On the value of mitigation and contingency strategies for managing supply chain disruption risks. *Management Sci.* **52**(5) 639–657.

Tu, Y, MO Ball, WS Jank. 2008. Estimating flight departure delay distribution: A statistical approach with long-term trend and short-term pattern. *J. Amer. Statist. Assoc.* **103**(481) 112–125.

Van Mieghem, JA. 2011. Risk management and operational hedging: An overview. P Kouvelis, L Dong, O Boyabatli, R Li, eds., *The Handbook of Integrated Risk Management in Global Supply Chains*. John Wiley & Sons Inc, Hoboken, NJ.

Wang, Y, W Gilland, B Tomlin. 2010. Mitigating supply risk: Dual sourcing or process improvement? *Manufacturing Service Oper. Management* **12**(3) 489–510.

Appendix A: Supplementary Material of Data and Results

A.1. Summary Statistics

Table A.1 provides summary statistics with predictors defined in Table 1.

Table A.1 Summary statistics

Dependent Variable						
	mean	std				
transport risk (hour)	-2.6	20.6				
Predictors						
Category Predictor						
	airline	route	airline-route	month	airline-leg2	airline-leg3
dimension	20	1336	588	7	20	16
Continuous Predictor						
	dev_{start} (day)	dur (day)	log(wgt) (kg)	log(pcs) (cbm)		
mean	-0.327	1.75	4.91	1.29		
std	0.648	1.30	2.4	1.43		

A.2. Model Parameter Estimation

Table A.2 shows the posterior mean and 95% probability interval of (selected) model parameters.

Table A.2 Posterior summaries of model parameters

Kernel Parameters					
μ_l $(l=1,2,\cdots,50)$	$\min(\mu_l)=-79.6,$	$\max(\mu_l)=76.01$			
$1/\sqrt{\phi_l}$ $(l=1,2,\cdots,50)$	$\min\left(1/\sqrt{\phi_l}\right)=0.72,$	$\max\left(1/\sqrt{\phi_l}\right)=84.4$			
Parameters in Weight γ					
Category Predictors					
θ_l^1 $(l=1,2,\cdots,49)$	$\min(\theta_l^1)=-10.9,$	$\max(\theta_l^1)=6.74$			
θ_a^2 $(a=1,2,\cdots,20)$					
θ_1^2 A1	θ_2^2 A2	θ_3^2 A3	θ_4^2 A4	θ_5^2 A5	
0	0.03	-5.27	5.15	3.09	
(0, 0)	(-0.40, 0.61)	(-5.86, -4.83)	(4.31, 6.11)	(2.89, 3.26)	
θ_6^2 A6	θ_7^2 A7	θ_8^2 A8	θ_9^2 A9	θ_{10}^2 A10	
1.16	8.53	2.54	-0.82	2.90	
(0.84, 1.53)	(8.19, 8.91)	(2.01, 2.98)	(-1.22, -0.40)	(2.23, 3.64)	
θ_{11}^2 A11	θ_{12}^2 A12	θ_{13}^2 A13	θ_{14}^2 A14	θ_{15}^2 A15	
-3.35	5.74	-2.96	2.74	-2.82	
(-4.02, -2.74)	(5.44, 5.97)	(-3.19, -2.67)	(2.27, 2.98)	(-3.26, -2.36)	
θ_{16}^2 A16	θ_{17}^2 A17	θ_{18}^2 A18	θ_{19}^2 A19	θ_{20}^2 A20	
4.95	-3.16	-5.36	6.23	-2.34	
(4.35, 5.50)	(-3.50, -2.76)	(-6.59, -4.41)	(5.79, 6.67)	(-2.58, -2.12)	
θ_{leg}^5 $(leg=2,3)$					
θ_2^5	θ_3^5				
-0.29	-0.34				
(-0.38, -0.21)	(-0.47, -0.21)				
Hyper-parameters					
$1/\sqrt{\epsilon^1}$	$1/\sqrt{\epsilon^2}$	$1/\sqrt{\epsilon^3}$	$1/\sqrt{\epsilon^4}$	$1/\sqrt{\epsilon^5}$	$1/\sqrt{\epsilon^6}$
4.86	3.39	6.26	7.02	0.64	0.74
(3.98, 5.93)	(2.62, 4.44)	(5.86, 6.63)	(6.46, 7.60)	(0.46, 0.90)	(0.51, 1.10)