Submitted to the Annals of Applied Statistics

A FORENSIC STATISTICAL ANALYSIS OF FRAUD IN THE FEDERAL FOOD STAMP PROGRAM

BY JONATHAN WOODY[1,a], ZHICONG ZHAO[1,b], ROBERT LUND,[2,d], AND TUNG-LUNG WU [1,c],

[1]*Department of Mathematics and Statistics, Mississippi State University,* [a]*jwoody@math.msstate.edu;* [b]*zz204@msstate.edu;* [c]*twu@math.msstate.edu*

[2]*Department of Statistics, University of California, Santa Cruz,* [d]*rolund@ucsc.edu*

This study develops methods to detect anomalous transactions linked with fraud in food stamp purchases through order statistics methods. The methods detect clusters in the order statistics of the transaction amounts that merit further scrutiny. Our techniques use scan statistics to determine when an excessive number of transactions occur (cluster), which is historically linked to fraud. A scoring paradigm is constructed that ranks the degree in which detected clusters and individual transactions are anomalous among approximately 250 million total transactions.

1. Introduction. This study examines Mississippi's food stamp records for anomalous transactions indicative of fraud. A Markovian relationship between successive order statistics is developed for independent and identically distributed (IID) data and used to statistically quantify clustering properties in store transaction records via scan statistic methods. The results illuminate some interesting features found around multiples of 65 dollars. This is the value of food coupon books of yesteryears and is often linked with fraud.

1.1. *SNAP program background.* The United States (US) government supports the Supplemental Nutritional Assistance Program (SNAP), frequently referred to as the food stamp program. The SNAP enables low-income families to purchase food, helping them meet basic nutritional needs. In the fiscal year (FY) 2018, approximately 40 million SNAP participants received an average benefit of about $245 per person per period or about $450 per household per period, with a total program cost of $65.3 billion dollars (Canning and Stacy, 2019).

The SNAP program history is narrated in Council et al. (2013). Early versions of the program distributed benefits via paper coupon books, with individual coupons called "food stamps". Coupons were intended to be used as a tax-free substitute for cash to purchase approved food and beverages. Figure 1 depicts a typical coupon book worth $65.00.

Beginning in the early 2000s, paper "food stamps" were replaced with Electronic Benefit Transfer (EBT) cards. SNAP benefits were thereafter deposited monthly to an account that can be accessed via EBT cards. EBT cards, which function similarly to debit cards, shift recipient benefits from a federal account to the SNAP retailer upon purchase of approved foods and beverages. The benefit recipient can purchase any amount (including less than the full retail price), tax free, of approved items. For example, a benefit recipient making a purchase of $100 worth of approved items could allocate $50 of SNAP benefits (tax free) and purchase the remaining $50 (plus tax) with their own cash or bank card. From the EBT card transactions, various state and federal agencies amassed large data sets.

SNAP fraud is relatively rare and can include fraud by households applying for benefits, application fraud by ineligible retailers, and fraud by state agencies (Cline, D.R. and Aussenberg R.A., 2018). The act of exchanging SNAP benefits for cash is also called SNAP

Keywords and phrases: Order statistics, Fraud detection, Scan statistics, Data science, SNAP trafficking.

Fig 1: A food stamp book containing $65 of coupons.

trafficking—this will be the only type of fraud considered this study. The United States Department of Agriculture (USDA) provided $7.5 million in the 2014 Farm Bill for states to create or improve technology systems to prevent or detect SNAP trafficking (Dean, 2016); thus, this is a potential new area of academic funding and interest beyond the state of Mississippi.

This study is the first detailed statistical analysis of an EBT transaction record. The data here contain all EBT transactions in Mississippi up to October of 2017 and is described further in the next section.

SNAP trafficking occurs in several manners. In one form, recipients exchange cash for their benefits at a discount from the vendor. For example, a vendor may debit an EBT card for $65 and give $50 in cash (Faulk, 2016). Another SNAP fraud scheme has beneficiaries selling or trading their benefits to other individuals rather than a vendor. For example, a recipient could exchange their benefits to a neighbor for cash, goods, or other services.

A summary of government estimated SNAP fraud rates is given in Appendix E of Wilson (2017); these figures are based on the Food and Nutrition Service's covert investigations. These investigations focus on retailers exhibiting suspicious behavior (Wilson, 2017). A summary of findings include: 1) trafficking diverted an estimated $1.1 billion annually from SNAP benefits; 2) approximately 1.5 percent of overall SNAP benefits were trafficked; and 3) approximately 11.8 percent of all authorized SNAP stores engaged in some form of trafficking.

1.2. *Statistical fraud detection techniques.* Many previous fraud detection methods appealed to Benford's Law, whereby the distribution of the leading digit of the transaction is scrutinized (Durtschi, Hillison and Pacini, 2004). However, Benford's Law requires the transaction amounts to span several orders of magnitude to realistically apply (Miller, 2015), which is not the case with food stamps — SNAP benefits are generally too meager for Benford's Law. A review of statistical fraud detection methods may be found in Bolton and Hand (2002).

Anomaly detection methods for fraud are surveyed in Chandola, Banerjee and Kumar (2009), and are well-developed in financial and cyber-security applications (Kou et al., 2004; Ahmed, Mahmood and Islam, 2016; Thiprungsri and Vasarhelyi, 2011; Liao et al., 2013). Anomaly detection is often viewed as a classification problem, with each transaction classified as either anomalous or normal (Aggarwal, 2014). Both supervised (Maes et al., 2002; Raj and Portia, 2011) and unsupervised learning methods (Abdallah, Maarof and Zainal, 2016; Hilas and Mastorocostas, 2008) have been applied to fraud classification problems in finance. Outlier methods (Aggarwal, 2015; Torgo, 2011; He et al., 2005; Ngai et al., 2011) have also

Fiscal Year	Transaction Count	FNS Count	Fiscal Year	Transaction Count	FNS Count
2002	1,850	251	2010	27,833,749	2936
2003	24,054	974	2011	30,553,070	3204
2004	39,461	1064	2012	32,574,197	3392
2005	78,969	1624	2013	34,032,660	3520
2006	171,360	2019	2014	31,804,535	3696
2007	458,461	2355	2015	32,229,687	3700
2008	2,028,258	2563	2016	29,067,111	3741
2009	16,444,289	2796	2017	13,038,192	3588
Total	250,379,903	7291			

TABLE 1
EBT transaction and vendor counts by fiscal year.

been used on fraud problems. Ekin et al. (2018) provide a comprehensive fraud assessment and a detailed review of outlier methods for medical data. Other related data mining-based methods to detect financial fraud can be found in Phua et al. (2010) and Al-Hashedi and Magalingam (2021). Unfortunately, anomaly and outlier methods are not applicable here as our clusters often contain many transactions that are neither anomalies or outliers.

In what ensues, we seek to identify clusters of transactions that are close to some fixed "price point". While the issue is tantamount to finding a mode in a probability density function, kernel density estimation techniques are not particularly useful (see Silverman (1986) and Good and Gaskins (1980) for mode identification techniques in kernel density estimates). Indeed, a typically identified modal region in a density estimate contains far too many prices to suggest anything about fraud. Phrased another way, typical fraud clusters are quite localized and will be smoothed away by any reasonably chosen density estimate bandwidth or histogram binwidth. Scan statistics are well-recognized as a powerful method to detect localized clusters in many applications (Konijn et al., 2013; Wu and Glaz, 2019). Liu and Zhang (2010) proposed a scan statistic-based method that can efficiently detect money laundering. Shao et al. (2021) develop a framework using non-parametric scan statistics to detect anomalous connected subgraphs, which can be indicative of fraud. This study develops a scan statistic approach that is capable of illuminating more localized features in the transactions.

The scan statistic will be applied to gaps in successive order statistics whose evolution is quantified via a Markovian relationship. This approach quantifies when observations cluster more than expected. We know of no other literature that uses order statistics methods for fraud detection.

The rest of this paper proceeds as follows. The next section discusses the EBT transaction data that drive this paper. Section 3 presents an exploratory data analysis of this data with rudimentary methods. Section 4 develops the order and scan statistic methods needed in our analysis. Section 5 presents a simulation study showing the efficacy of the methods on synthetic data. Results for our Mississippi transactions are presented in Section 6. Comments and conclusions conclude the article in Section 7.

2. The Data. Our data contain all historical EBT transactions in Mississippi from 2002 through October of 2017. The number of transactions by year are listed in Table (1). The large increase in transactions from FY 2008 to FY 2009 was driven by two events. First, the Farm Bill of 2008 moved all benefits to EBT cards and ended paper food stamps. Second, a severe recession occurred in 2008, causing a rise in SNAP benefit applications.

Seven fields of information come with each SNAP transaction; these are listed in Table (2). Field 1 is the case number. Each household receives a unique case number every time they enroll for SNAP benefits. There is only one case number per household; hence, one cannot

differentiate between multiple purchasers within a household. Field 2 is the date of purchase in the format month/day/year. Field 3 demarcates the time of day of the transaction and Field 4 the transaction amount. Field 5 contains the merchant's name. Chain stores can have multiple locations with the same name. This field is often subject to mistyped entries. Field 6 is the town where the transaction took place; this field also has frequent typographical errors. Field 7 shows the FNS number: each merchant location has a unique FNS number identifying their store.

Field 1	**Field 2**	**Field 3**	**Field 4**	**Field 5**	**Field 6**	**Field 7**
Case Number	Date	Time	Amount	Merchant Name	Town	FNS Number

TABLE 2
The seven fields of data associated with each EBT transaction.

Some aspects of the SNAP program in Mississippi are now clarified. A fiscal year (FY) is defined to start on October 1 of the previous calendar year and ends on September 30. Our focus is limited to Mississippi residents making transactions at SNAP accepting Mississippi vendors during FY2008 - FY2016. We also include partial FY 2017 transactions, with the data stopping on 3/29/2017. Retailers accepting EBT transactions will be called vendors. Each household receiving benefits is given a unique case number. Should a household cease receiving SNAP benefits and return to the program at a later date, they receive a new case number.

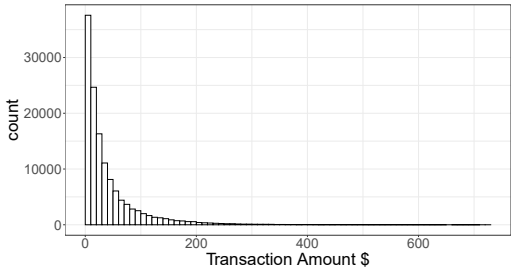
The number of EBT accepting vendors changes with time since stores may begin or stop accepting EBT transactions. This happens when new EBT accepting vendors are created or go out of business. In this study, each physical store is considered a distinct EBT accepting vendor. Therefore, a chain having several stores with the same name are considered separately. Table 1 provides further information during each FY in our study.

3. Exploratory Data Analysis. This section illuminates some transactions that have been previously associated with fraud (SNAP trafficking). While irregularities have been reported around price points that are integer multiples of $65, price point irregularities may appear at other unsuspected transaction amounts. For example, transactions at integer multiples of $10 (exactly) are observed at vendors that serve gasoline; this coincides with non-fraudulent credit card purchases of £10 petroleum observed in England (Hand et al., 2000).

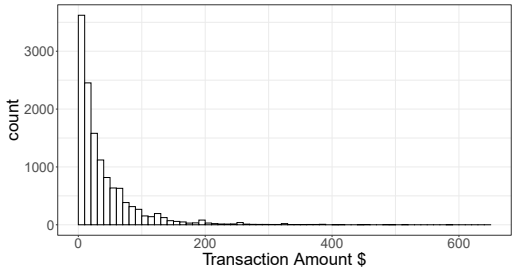
To explore the data, histograms of all EBT transactions during each fiscal year were produced for each vendor. If a vendor only accepted EBT transactions during part of a FY, then only transactions during that portion of the year are considered; there is no effort to rescale any transaction frequencies to a full FY.

Figure 2 depicts transaction histograms from two vendors during FY 2015, whom we dub Vendors A and B. These two vendors will be used for case studies throughout this paper and are selected solely to demonstrate some commonly occurring transaction patterns. Vendors A and Vendor B were selected as archetypal examples of no clustering (Vendor A), and "variance-free" clustering at integer multiples of a $65 price point (Vendor B). In Section 6, an example of price point clustering with variance is presented (this merchant is dubbed Vendor C).

Vendor A experienced roughly 10 times as many transactions as Vendor B. While the two histograms exhibit a similar structure, the businesses are dissimilar: Vendor A is a large chain grocery store, while Vendor B is a niche vendor operating only one store. The histogram binwidth used was $10, smaller than the default selected by R.



(a) Vendor A EBT transactions with $10 bins.

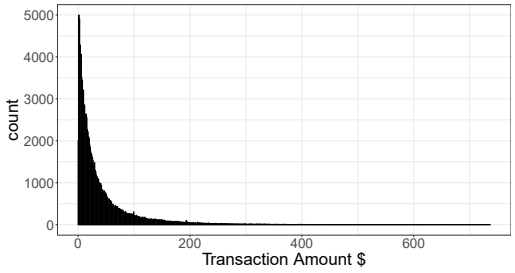


(b) Vendor B EBT transactions with $10 bins.

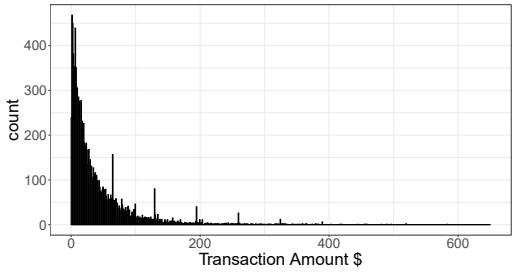
Fig 2

Differences emerge when one decreases the binwidths. Figure 3 shows histograms of the above transactions when the binwidth is decreased to $1.00. Note that clusters now appear in Vendor B's histogram at integer multiples of $65.00, while Vendor A's histogram remains "spikeless". These spikes are not due to expenditure of full coupon book amounts — paper food stamps were replaced by the EBT system in 2008.

The binwidths for the histograms are selected to demonstrate that hidden patterns may exist in the transactions — some sleuthing may be warranted. While optimal binwidth selection criteria exist (Scott, 1979; Wand, 1997), histograms with differing binwidths are presented here for feel only. An empirical cumulative distribution function (ECDF) could also be used to visually identify density spikes; however, many of the spikes encountered are so localized that they will not be visually evident in an ECDF.



Vendor A EBT transactions with $1 bins.



Vendor B EBT transactions with $1 bins.

Fig 3

The vast majority of vendors in our data do not exhibit anomalous clusters above prices for a few popular individual items (such as a sandwich or drink). Vendor A is an example of such a vendor. Of those showing anomalous clusters, patterns vary among vendors. Vendor B has a transaction distribution with clusters at integer multiples of $65 with little or no variance (that is, the transactions are at exact multiples of $65). State investigators believe that some vendors try to hide fraud by adding a small amount (more or less) to the $65 benchmark (see the discussion of Vendor C in Section 6). Additionally, the price point where the clusters occur vary by vendor and time. Overall, the statistical challenge remains the same: identify clustering of price points in the data. This will be tackled with order and scan statistics methods in the next section.

Before continuing, we comment that identified clusters need not be associated with fraud. Indeed, the following are potential explanations of clusters not linked to fraud.

1. Individual SNAP recipients have learned to spend monthly benefits by the denominations in the old coupon books.
2. Merchants have targeted SNAP users with products priced at the old coupon book denominations (although we see no evidence of this).
3. Other reasons not identified here.

Likewise, the following cluster explanations are linked to fraud.

1. SNAP recipients exchange SNAP benefits for other individuals in exchange for cash or goods in denominations of the old coupon books.
2. SNAP recipients exchange SNAP benefits directly with vendors in exchange for cash or non-SNAP eligible goods in denominations of the old coupon books.
3. Other reasons not identified here.

4. Methods. We now develop the methods used to identify local transaction clusters embedded in a background distribution. We do not know the background distribution *a priori*, nor the price points where the clusters appear. Our task is to construct a fully automated cluster detection system that works for all 7,291 vendors, which collectively have a variety of different transaction price distributions. Our end goal is to flag any suspicious transactions and to rank order these in terms of their suspiciousness.

Our methods will examine the order statistics of the price transactions, focusing on the gaps between these transactions. Order statistics are studied in David and Nagaraja (2004); Arnold, Balakrishnan and Nagaraja (2008); Ahsanullah, Nevzorov and Shakil (2013) and connected to Poisson processes via conditional uniformity of arrivals in Karlin and Taylor (1981); Feller (1966); Liberman (1985); Feigin (1979), for example. From the order statistics, we construct a sequence of Bernoulli trials that is used to assess clustering properties of the transactions. Our main cluster identification tool is the scan statistic, which can capably identify small intense clusters (local departures). Flexibility of the methods is important as the seven thousand plus vendors have a plethora of different distributions.

4.1. *Data preparation.* Let $\{X_i^*\}_{i=1}^n$ denote the assumed IID transaction amounts for $1 \leq i \leq n$ at a specified vendor in a given FY. Here, n is the total number of transactions and X_i^* is the i^{th} transaction amount. The data is discrete, measured in dollars and cents. A "jitter" U_i in the form of a uniform random variable on (-1,0)¢ is added to X_i^* to convert the transaction to a continuous random variable, analogous to Nagler (2018), with minimal effect on any results: $X_i = X_i^* + U_i$. Of course, $\{X_i\}_{i=1}^n$ is also IID and no two X_is are the same with probability one.

Visual inspection of the Section 3 histograms indicate that the transaction amounts are reasonably approximated by the Gamma density

$$f_X(x) = \frac{x^{\alpha-1}e^{-\beta x}\beta^\alpha}{\Gamma(\alpha)}, x > 0,$$

where $\alpha, \beta > 0$. However, this choice of distribution will only be used to map the transactions to (0,1); other continuous distributions can be used. In particular, we will work with the cumulative distribution function (CDF) transformed data $V_i = \hat{F}_X(X_i)$ for $i = 1, \ldots, n$, where $\hat{F}_X$ is the Gamma CDF evaluated at X_i, parameterized by the MLE estimates of $\hat{\alpha}$ and $\hat{\beta}$. By construction, $\{V_i\}_{i=1}^n$ are IID continuous random variables with support set $[0, 1]$. Let F_V and f_V denote the CDF and probability density function of the data. If the data is indeed Gamma distributed, the V_is are IID Uniform[0,1] (up to error induced by estimation of the Gamma parameters). However, in truth, the data may be non-Gamma and the V_i need not be precisely uniformly distributed on [0,1]; nonetheless, this is a convenient way to proceed since order statistics of uniform[0,1] variables are well understood. Later, we will impose regularity assumptions on f_V.

4.2. *Analyzing the gaps.* Let $0 := V_{(0)} < V_{(1)} < \cdots < V_{(n)} < V_{(n+1)} := 1$ be the order statistics of $\{V_i\}_{i=1}^n$, the inequalities being strict since each V_i is unique with probability one. Let

$$G_i := V_{(i)} - V_{(i-1)} \tag{1}$$

be the size of the ith gap. While the gaps $G_1, \ldots, G_{n+1}$ are known to be exchangeable, G_i and G_j are not necessarily independent or identical in distribution when $i \neq j$ (Arnold, Balakrishnan and Nagaraja (2008)); in fact, the dependence $G_1 + \ldots + G_{n+1} = 1$ holds. If $V_{(i-1)} = v_{(i-1)}$, then we know that $i-2$ values in the sample are less than $v_{(i-1)}$, $n-i+1$ sample values exceed $v_{(i-1)}$, and one observation exactly equals $v_{(i-1)}$. For the $n-i+1$ sample values that exceed $v_{(i-1)}$, we do not have any additional information about them — just that they exceed $v_{(i-1)}$. Hence, conditional on $V_{(i-1)}$, these exceeding values can be regarded as being drawn from the probability density function

$$\frac{f_V(v)}{1 - F_V(V_{(i-1)})}, \quad V_{(i-1)} < v < 1. \tag{2}$$

The upshot is that the Markov-type relationship

$$V_{(i)} \stackrel{\mathcal{D}}{=} V_{(i-1)} + \min_{1 \leq i \leq n-i+1} (R_1, \ldots, R_{n-i+1}), \tag{3}$$

governs successive order statistics, where the R_js are IID and have the distributional form $R_j = M_j - V_{(i-1)}$ and the M_js are IID, have the distribution in (2), and most importantly, are independent of $\{V_{(k)}\}_{k=1}^{i-1}$. This relationship essentially drives our work.

An implication of the above is that the distribution of G_i given $V_{(1)}, \ldots, V_{(i-1)}$ only depends on $V_{(i-1)}$ and not on any previous order statistic (or gaps). Hence, for $t_i \in [0,1]$,

$$P\left(G_i \leq t_i \middle| V_{(1)}, \ldots, V_{(i-1)}\right) = P\left(G_i \leq t_i \middle| V_{(i-1)}\right). \tag{4}$$

Now apply the chaining relationship

$$P(A_1 \cap \ldots \cap A_n) = P(A_1)P(A_2|A_1)\cdots P(A_n|A_{n-1}, \ldots, A_1)$$

for any events $A_1, \ldots, A_n$ and (3) and invoke the conditional independence in (4) to get

$$P(G_1 \leq t_1 \cap \cdots \cap G_n \leq t_n) = \prod_{i=1}^{n} P_{V_{(i-1)}}(G_i \leq t_i). \tag{5}$$

Here, the Markov style notation for probabilities $P_{v_{(i-1)}}(G_i \leq t_i) := P(G_i \leq t_i | V_{(i-1)} = v_{(i-1)})$ has been adopted. We also use this notation for expectations; for example, $E_{V_{(i-1)}}[G_i] := E[G_i|V_{(i-1)}]$.

To quantify where the $V_{(i)}$s cluster (and hence the X_is), we will be interested in when many small gaps occur in close proximity. To develop a scan statistic approach to quantify this, set

$$Y_i(\theta) = 1_{\{G_i \leq \theta E_{V_{(i-1)}}[G_i]\}}, \tag{6}$$

for a tuning parameter $\theta > 0$ that is utilized in our ranking paradigm and discussed further below. The above arguments imply that conditional on $V_{(i-1)}$, these indicator variables are independent for every n, viz.,

$$P(Y_1 = j_1, Y_2 = j_2, \ldots, Y_n = j_n) = \prod_{k=1}^{n} P_{V_{(k-1)}}(Y_k = j_k) \tag{7}$$

for $j_1, \ldots, j_n$ taking values in $\{0, 1\}$. Note that $Y_i(\theta) = 1$ indicates that $V_{(i-1)}$ and $V_{(i)}$ are closer together than expected.

While the indicators are conditionally independent, they do not necessarily have a common success probability (the distribution of the gap sizes depends on n in general). Indeed, our next objective is to show that as $n \to \infty$, the success probability of these indicators essentially converges to $1 - e^{-\theta}$ when they are suitably far away from the support set edge at unity.

Returning to the method, since $f_v(v) > 0$ for all $v \in (0, 1)$, as $n \to \infty$, $v_{(i-1)} \downarrow 0$ for any fixed i. Define a pth quantile as

$$q(p) = \inf\{v \in (0, 1) : F_V(v) \geq p\} \tag{8}$$

and the index where data becomes larger than a fixed x as

$$\hat{i}_n(x) = \min\{i : v_{(i)} > x\}, \tag{9}$$

which depends on the sample size n. Since $f_V(v) > 0$ for all $v \in (0, 1)$, $v_{(\hat{i}_n(q(p)))} \downarrow q(p)$ and $v_{(\hat{i}_n(q(p))-1)} \uparrow q(p)$ as $n \to \infty$ for any $p \in (0, 1)$. When the index n and $q(p)$ are unimportant, future notations will suppress these quantities for ease of exposition.

To show that the indicators in (6) converge to trials with the common success probability $1 - e^{-\theta}$, note that since $f(v) > 0$ for all $v \in (0, 1)$, $n - \hat{i}_n \to \infty$ for any quantile $q(p) < 1$.

To prove our main result, we first present a useful Lemma. Let $h(v) = f_V(v)/[1 - F_V(v)]$ be the hazard rate function (HRF) of the V_is and set $H(v) = \int_0^v h(t)dt$ as the cumulative HRF. The following results hold for any fixed i. The CDF and PDFs of the V_is are well known to be

$$F_V(v) = 1 - \exp\{-H(v)\}, \quad f_V(v) = h(v)\exp\{-H(v)\}. \tag{10}$$

Conditional on $V_{(i-1)}$ for a fixed i, (3) shows that the support set for the R_js is $(0, 1 - V_{(i-1)}]$ for $j = 1, \ldots, n - i + 1$. For this i, let $F_{R_j|V_{(i-1)}}$ denote the conditional CDF of one such R_j. We now have

$$F_{R_j|V_{(i-1)}}(t) = \frac{F_V(V_{(i-1)} + t) - F_V(V_{(i-1)})}{1 - F_V(V_{(i-1)})} = 1 - \exp\{-(H(V_{(i-1)} + t) - H(V_{(i-1)}))\}.$$

The corresponding probability density function is hence

$$f_{R_j|V_{(i-1)}}(t) := h(V_{(i-1)} + t)\exp\{-(H(V_{(i-1)} + t) - H(V_{(i-1)}))\}. \tag{11}$$

Watson and Leadbetter (1964) state that

$$E\left[\frac{1}{h(V_{(i)})}\right] = (n - i + 1)E[G_i] \tag{12}$$

for $i = 0, \ldots, n - 1$. A conditional version of (12) will prove useful in our ensuing arguments.

LEMMA 4.1. *Let $\{V_i\}$ be IID continuous random variables on $(0, 1)$. Then*

$$E_{V_{(i-1)}}\left[\frac{1}{h(V_{(i)})}\right] = (n - i + 1)E_{V_{(i-1)}}[G_i]. \tag{13}$$

The proof of Lemma (4.1) is given in the Appendix. With Lemma 4.1, our main result can now be stated.

THEOREM 4.2. *Let $\{V_i\}_{i=1}^n$ be IID continuous random variables on $(0,1)$ with a Lipschitz continuous density function f_V satisfying $f_V(v) > c$ for some constant $c > 0$ for all $v \in (0,1)$. If $\{\hat{i}_n\}$ is any sequence of indices such that $\hat{i}_n \to \infty$, and $n - \hat{i}_n \to \infty$, then $P_{V_{(\hat{i}_n-1)}}(Y_{\hat{i}_n}(\theta) = 1) \to 1 - e^{-\theta}$.*

The proof of Theorem (4.2) is given in our Appendix.

Note that the success probabilities of the Bernoulli $Y_i(\theta)$ in (6) depend on the conditional expected gap size $E_{V_{(\hat{i}_n-1)}}[G_{\hat{i}_n}] = E_{V_{(\hat{i}_n-1)}}\left[\frac{1}{h(V_{(\hat{i}_n)})}\right]/(n - \hat{i}_n + 1)$; that is, the expected gap sizes depend on $v_{(\hat{i}_n-1)}$'s since h is a function of $v_{(\hat{i}_n-1)}$.

We now introduce a useful and appealing approximation. For each $\hat{i}_n$,

$$E_{V_{(\hat{i}_n-1)}}[G_{\hat{i}_n}] \approx \frac{1}{(n+1)f_V(V_{(\hat{i}_n-1)})}. \tag{14}$$

The approximation is justified in the Appendix. The interpretation of (14) is intuitive: as the sample size increases, expected gap sizes decrease. Likewise, gaps are smaller at places where the density is larger since more observations are likely to occur nearby.

4.3. *Estimation of the background density.* This subsection estimates $f_V(\cdot)$, which is needed to compute our scan statistics. The density $f_V(\cdot)$ for the CDF transformed transactions should be free of spikes if there is no fraud. In practice, this density is usually close to the uniform [0,1] density since the transactions were transformed with a parametric-fitted Gamma CDF. The method will prove robust to poorly fitting Gamma CDFs however, as will be demonstrated in the case studies below (see Vendor C).

To quantify departures from uniformity, let $0 < p_1 < p_2 < \ldots < p_m < 1$ be ordered points in $(0,1)$ with the boundary settings $p_0 = 0$ and $p_{m+1} = 1$. Let $\Pi = \{p_0, p_1, \ldots, p_m, p_{m+1}\}$ be a partition of [0,1]. We assume that f_V belongs to the class of density functions formed by linear interpolation of points in Π, viz.,

$$f_V(v) = a_i v + b_i, \quad p_{i-1} < v \le p_i. \tag{15}$$

There are restrictions to this class. First, we require that a_i and b_i are such that $f_V(v)$ is continuous in v. Second, to have a non-negative (legitimate) density, we assume that $a_i v + b_i > 0$ over $v \in (p_{i-1}, p_i]$. There are no nice forms for the likelihood estimators of a_i and b_i for such a density. Because of this, we employ an ad-hoc but simple approach to estimate $f_V(\cdot)$ using kernel density estimation. Other approaches were examined, but overall results did not seem to greatly change.

In our numerical work, the partition

$$\Pi = \{0.00, 0.05, 0.10, \ldots, 0.95, 1.00\} \tag{16}$$

was deemed appropriate after exploratory analysis. Density estimates of $f_V(v)$ are estimated for each point $v = p_i$ via a kernel approach. To handle v near the boundaries of 0 and 1, we reflect the sampled values $\{V_1, \ldots, V_n\}$ about zero and unity as in Silverman (1986). A Gaussian kernel is then employed without undue edge effects. While bandwidth selection rules in Silverman (1986) recommended a bandwidth proportional to $n^{-1/5}$, we prefer an oversmoothed version of the density to an undersmoothed one. Indeed, this is why we prefer the class of density functions depicted in Equation (15). After much exploratory analysis, we found $h = n^{-1/2}$ ideal for our needs. Admittedly, one could argue for adjustments to this on a case-by-case basis.

Next, the density estimated values $\hat{f}_V(p_i)$ are rescaled into a proper density function by numerically imposing that

$$1 = \int_0^1 \hat{f}_V(v)dv = \sum_{i=0}^{m} (1/2)\left[\hat{f}_V(p_{i+1}) - \hat{f}_V(p_i)\right](p_{i+1} - p_i). \tag{17}$$

4.4. *The scan statistic.* This subsection develops a scan statistic capable of detecting clusters. To develop the statistic, assume that $\{Y_i(\theta)\}_{i=1}^n$ are IID Bernoulli trials with success probability $1 - e^{-\theta}$. For a window length $r \geq 1$ and $i \in \{1, \ldots, n-r\}$, define

$$S_n(r, i, \theta) = \sum_{k=i}^{i+r-1} Y_k(\theta). \tag{18}$$

An unconditional discrete scan statistic with window size r is the maximum

$$S_n(r, \theta) = \max_{1 \leq i \leq n-r} S_n(r, i, \theta). \tag{19}$$

The scan statistic in (19) is statistically complex to quantify as it involves the maximum of highly dependent rolling windows. A ubiquitous challenge with scan statistics involves computations of its distributions. The exact distribution of $S_n(r, \theta)$ is given in Fu (2001) and is obtained from a finite Markov chain embedding technique. Unfortunately, exact probabilities have only been numerically computed for $r \leq 35$ as the problem is NP-hard. Developing algorithms to efficiently compute the exact distributions of scan statistics is an open problem.

Naus (1982) proposed an approximation of the scan statistic's distribution based on its Markov-type structure. In many applications, researchers simply resort to Monte Carlo simulations to estimate the distribution of the scan statistic. For large window sizes r, Haiman (2007) proposed an accurate approximation that treats discrete scan statistics as extremes of one-dependent stationary sequences. Haiman's approximation for scan statistic probabilities is used here.

Atypical sequences of successes will produce abnormally high $S_n(r, i, \theta)$ in (18). For a given significance level α, a cluster is signaled for any $i \in \{1, 2, \ldots, n-r+1\}$ whenever $S_n(r, i, \theta)$ exceeds an α-level threshold T_α. The α-level threshold is defined as

$$T_\alpha = \min_{T=0,1,\ldots,r} \{T : \ P(S_n(r, \theta) > T) \leq \alpha \}. \tag{20}$$

There is a slight nuance regarding the α-level of the test since the scan stat in (19) is discrete: the true α-level of the test with threshold T_α is

$$\alpha^* = P(S_n(r, \theta) > T_\alpha), \tag{21}$$

where $0 \leq \alpha^* \leq \alpha$.

Our exploratory analysis revealed potential fraudulent transactions at price points that are integral multiples of \$65 (among other dollar amounts). To detect this type of fraud, we examine the number of transactions between each dollar amount and take the maximum over all dollars. Specifically, our window length used is

$$r = \max_{d=0,1,2,\ldots} (C_d - E[C_d]), \tag{22}$$

where C_d is the number of transactions between d and $d+1$ dollars (the added jitter to each transaction will not change the $[d, d+1)$ interval that it lies in) and $E[C_d]$ is computed from a total of n transactions and a $\hat{f}_X$ estimated from the MLE fit of the Gamma distribution. Window size selection in scan statistics is usually arbitrary as true cluster sizes are typically unknown. Minimizing the expected time until first detection or maximizing test power are alternative ways of selecting r. One can certainly try any of these, but the above procedure works well on this data set.

4.5. *Effects of CDF Estimation.* In practice, the estimated version of F_X may induce dependencies into $\{Y_i(\theta)\}$. The estimated versions of the V_is are

$$\hat{V}_i = \hat{F}_X(X_i),$$

where $\hat{F}_X$ is the MLE estimated CDF of the transaction distribution. Let $(\hat{\alpha}, \hat{\beta})'$ be estimators of the Gamma parameters from a sample of size n. Suppose that these estimators are nice and regular so that

$$\sup_x |\hat{F}_X(x) - F_X(x)| \longrightarrow 0$$

almost surely as $n \to \infty$, where $\hat{F}_X(x)$ is the estimated F_X from $\{X_i\}_{i=1}^n$. One can always use the empirical CDF and quote the convergence in the classical Glivenko-Cantelli Theorem if desired.

From this convergence, it can be shown that as $n \to \infty$, our methods still hold for $\{\hat{V}_i\}$. The same is true for $\{\hat{G}_i\}, \{\hat{E}_i\}$, and $\{\hat{Y}_i(\theta)\}$, which are constructed by replacing V_i by $\hat{V}_i$ in Equations (1) and (6) respectively. We do not prove this here, but mention that our argument requires showing that

$$\lim_{n \longrightarrow \infty} nE[|\hat{G}_i - G_i|] = 0,$$

which can be done through dominated convergence without additional assumptions on the density.

On a more practical level, simulations in Section 5 demonstrates that $\hat{E}_i \to E_i$ as $n \to \infty$ for most i where $V_{(i)}$ is sufficiently far from zero and unity. This implies that $\{\hat{Y}_i(\theta)\}$ behaves asymptotically as needed.

To consider the practical implications of the asymptotic results on finite sample sizes, a simulation study is performed in Section 5 for sample sizes of magnitude encountered with our data. The results show that $\{\hat{Y}_i(\theta)\}$ is approximately distributed according to Theorem 4.2. For this reason, we eschew "adding hats in the notation" going forward.

4.6. *Ranking flagged clusters.* This subsection presents a method to rank the degree to which flagged clusters are anomalous. Our task here is to construct a depth/metric that will assign a score to each flagged cluster.

We begin by aggregating all transactions for each vendor and each FY. Transactions are segmented by FY since the SNAP funds dispersed may change with the FY. Clusters at the maximum benefit amount are occasionally detected — and this maximum benefit changes from year to year. Hence, each FY is analyzed separately.

Consider a vendor FY where one or more clusters are flagged. Frequently, multiple clusters are flagged for a given vendor FY. Let m denote the number of clusters detected for this vendor FY. In the ensuing cluster rankings, we back transform to the original data in dollars, working with the transaction amounts $X_{(i)} = \hat{F}_X^{-1}(V_{(i)})$.

For a given θ and decision threshold η, the starting point of the first cluster is

$$\tau_1 := \min_{1 \le i \le n-r+2} \{i : S_n(r, i, \theta) > \eta\}, \tag{23}$$

and the ending point of the first cluster is

$$\kappa_1 := \min_{\tau_1 < i \le n-r+2} \{i : S_n(r, j, \theta) \le \eta, \text{ for all } j = i, i+1, \ldots, i+r-2\}. \tag{24}$$

Note that $S_n(r, i, \theta) > \eta$ for all $i \in \{\tau_1, \tau_1 + 1, \ldots, \kappa_1 - 1\}$. Additionally, $S_n(r, i, \theta) \le \eta$ for $i = \tau_1 - 1$ and $i = \kappa_1, \kappa_1 + 1, \ldots, \kappa_1 + r - 2$. Therefore, the first cluster contains the transactions $\{X_{(\tau_1)}, X_{(\tau_1+1)}, \ldots, X_{(\kappa_1+r-2)}\}$.

Continuing in this fashion, additional clusters are defined as subsequent order-index i sojourns, where $S_n(r,i,\theta) > \eta$. Specifically, the jth cluster has the beginning and ending indices

$$\tau_{j+1} := \min_{\kappa_j \le i \le n} \{i : S_n(r,i,\theta) > \eta\}, \tag{25}$$

and

$$\kappa_{j+1} := \min_{\tau_{j+1} < i \le n} \{i : S_n(r,i,\theta) \le \eta\}. \tag{26}$$

The jth cluster contains the transactions $\{X_{(\tau_j)}, X_{(\tau_j+1)}, \ldots, X_{(\kappa_j+r-2)}\}$ for $j = 2, \ldots, m$.

We now discuss the parameter θ. As $\theta \downarrow 0$, the success probability of $Y_i(\theta)$, which is $1 - e^{-\theta}$ asymptotically, tends monotonically down to zero. More clusters are signaled with a higher θ for a fixed threshold η.

Let T_α be the critical value of $S_n(r,i,\theta)$ when $\theta = 1$. For any transaction i, we define a minimum θ where the transaction is flagged as potentially anomalous, denoted by $\theta_{\min}$, to be the smallest θ having $S_n(r,i,\theta) > T_\alpha$. Here, T_α is held constant, set according to $\theta = 1$.

Conversations with Mississippi SNAP program officials indicate that some vendors try to mask illicit transactions by adding some small random amounts to the transactions. This avoids identical transaction amounts, which would be easier to identify. Because this sort of tactic involves clerks and is hence more nefarious, a cluster that exhibits a range/variance in its transactions is deemed more anomalous than a cluster with tightly packed transactions. Clusters with higher transaction amounts are also deemed more anomalous.

We define the depth of transaction $X_{(i)}$ as $D(i) = \theta_{\max} - \theta_{\min}(i)$ should this transaction be in a significant cluster as measured, where $\theta_{\max} = 1/2$ (presumption of no cluster mandates a conservative value here). A zero depth is assigned to the transaction otherwise. Our rank for cluster j incorporates the range of the transactions in the cluster and the depths via

$$R_j := \left(\sum_{i=\tau_j}^{\kappa_j - 1} X_{(i)} \right) \times \left(1/2 \sum_{i=\tau_j}^{\kappa_j - 1} [X_{(i+1)} - X_{(i)}][D(i) + D(i+1)] \right). \tag{27}$$

Intuition for this ranking method is included in the discussion below Figure (12), but note that the rank R_j has the following properties:

1. higher R_js are more anomalous,
2. clusters with larger sum totals are more anomalous,
3. clusters that are more intense (as measured by $D(i)$) are more anomalous,
4. clusters with a larger range (as measured by $X_{(\kappa_j - 1)} - X_{(\tau_j)}$) are more anomalous.

Assume that the procedure flags m clusters for a given vendor FY. Let $R_{(1)} < R_{(2)} < \cdots < R_{(m)}$ denote the ordered rankings, from least to most anomalous, of all flagged clusters (the inequalities are strict with probability one). When clusters are detected at a given vendor, the vendor is assigned the score $R_{(m)}$; otherwise, the vendor receives a zero score. Let $C_{(j)}$ denote the set of transactions having rank $R_{(j)}$ for $j \in \{1, 2, \ldots, m\}$. The most anomalous cluster is $R_{(m)}$ and has the transactions $C_{(m)}$.

Let $\tau_{(j)}$ and $\kappa_{(j)} + r - 2$ denote the starting and ending index of $C_{(j)}$ for $j \in \{1, 2, \ldots, m\}$. To rank individual transactions, first each transaction in $C_{(m)}$ is ranked from largest to smallest: $\{X_{(\kappa_{(m)}+r-2)}, \ldots, X_{(\tau_{(m)})}\}$. Next, the set of transactions $C_{(m-1)}$ are ordered analogously and appended to the right of $C_{(m)}$. This rank ordering continues for all clustered transactions until all flagged transactions for a given vendor are ranked as

$$\{X_{(\kappa_{(m)}+r-2)}, \ldots, X_{(\tau_{(m)})}, X_{(\kappa_{(m-1)}+r-2)}, \ldots, X_{(\tau_{(m-1)})}, \ldots, X_{(\kappa_{(1)}+r-2)}, \ldots, X_{(\tau_{(1)})}\}.$$

Non-flagged transactions are then discarded.

The methods were applied to all 7,291 vendors in this study, giving a rank to all 250,379,903 transactions; results are discussed in Sections 6 and 7. Note that the ranking system ranks vendors (via $R_{(m)}$) and transactions occurring at a given vendor, but does not directly compare transactions between two distinct vendors.

5. A Simulation Study. This section presents a simulation study of our cluster detection procedure. First, the procedure is tested when no clusters are present and f_V is known. Next, an assessment of the independence of $\{\hat{Y}_i(\theta)\}$ is considered when the transaction distribution is estimated, augmenting the discussion in Section 4.5. Next, the entire cluster detection procedure is studied under several scenarios to quantify Type I and II errors. Here, the true density f_V is assumed unknown and $\hat{f}_V$ is computed as described in Subsection 4.3. Power and error rates are assessed for a variety of window widths r. Throughout, random variables supported on [0,1] were generated via CDF-transformed observations.

5.1. *Simulation A.* First, the method is applied to the case where the density of $\{X_i\}$ is assumed known; this is equivalent to letting $\{V_i\}_{i=1}^n$ be IID Uniform[0,1] random variables. For this simulation, we use $n = 4,000$ and set $r = 30$. For $\alpha = 0.05$, the selection criteria in (20) yields $T_\alpha = 28$, which by (21), corresponds to an $\alpha^* = 0.024$ level test. One hundred independent replications of $\{V_i\}_{i=1}^n$ were generated. In each replication, the procedure of Section 4 is applied with $r = 30$ and $\theta_{\max} = 1$. The scan statistic $S_n(r, i, \theta)$ is computed for each i and compared to the threshold T_α to make decisions.

Table 3 depicts the number of flagged clusters over 100 independent runs. In this case, three erroneous clusters are detected (Type I errors); this is almost as expected for a level 0.024 test. Overall, the procedure appears to work quite well.

5.2. *Simulation B.* To investigate the asymptotic independence of $\{\hat{Y}_i(\theta)\}$ when the CDF is estimated, transactions from a Gamma(α, β) distribution, where $\alpha = 10$ and $\beta = 10$, were sampled. Three samples sizes — $n = 1,000$, $n = 10,000$, and $n = 100,000$ — were considered and $\{\hat{V}_i\}_{i=1}^n$ was calculated for each simulation run and sample size n. The estimates $\hat{\alpha}$ and $\hat{\beta}$ are obtained via the egamma function of the R EnvStats package. Note that if $E_i = \hat{E}_i$, then $Y_i(\theta) = \hat{Y}_i(\theta)$. Therefore, we study the ratio $E_i/\hat{E}_i$ to nullify the size effects of the E_is. The three panel graphic in Figure 4 displays results. One hundred independent samples were drawn for each n considered, the E_is and $\hat{E}_i$s were constructed, and the sample path $E_i/\hat{E}_i$ for each i was plotted for all 100 runs. When $n = 1,000$, over the index range of 25 to 975, this ratio is approximately unity — uniformly in the 100 runs. Hence, even for this smallest sample size, the independence of $\hat{E}_i$ in i seems reasonable. The cases $n = 10,000$ and $n = 100,000$ show an even better alignment. Thus, the assumption of independent $\{\hat{Y}_i(\theta)\}$ at the quantiles analyzed seems roughly valid for the vendors considered in this study.

5.3. *Simulation C.* Our next simulation, which studies what happens when the sample comes from a non-gamma family of distributions, considers non-uniformly distributed data with no clusters. In this case (and those that follow), we first simulate an IID sequence $\{W_i\}$ of non-uniform data on [0,1] when $n = 4,000$. In particular, the W_is are generated from the sinusoidal density

$$f_W(w) = 1 - 0.30\sin(2\pi w), \quad 0 < w < 1. \tag{28}$$

An inverse gamma CDF transform is then applied to each W_i using the parameters $\alpha = 2$ and $\beta = 50$: $X_i = F_X^{-1}(W_i)$, which leaves $\{X_i\}$ non-gamma distributed over $(0, \infty)$. Next, we estimate α and β from the sample $\{X_i\}_{i=1}^n$ and work with the CDF transformed data

$$\hat{V}_i = \hat{F}_X(X_i),$$

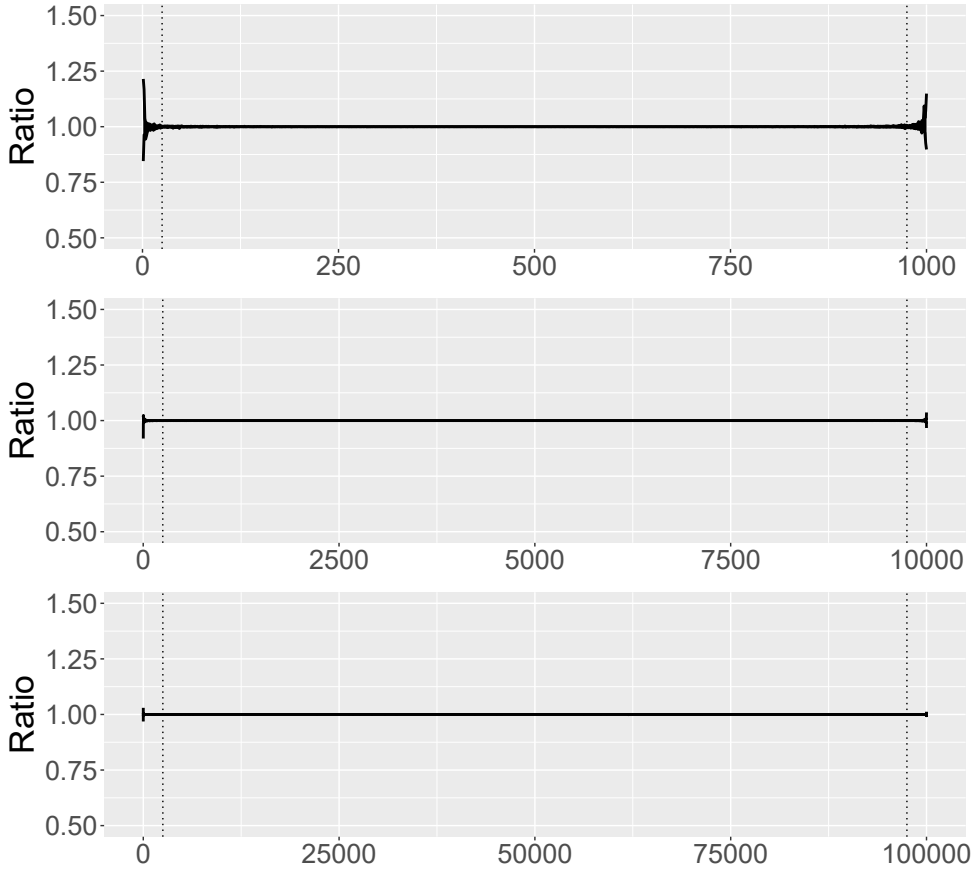


Fig 4: Effects of estimating the CDF. Here, 100 samples of size $1,000$ (top panel), $10,000$ (middle panel) and $100,000$ (bottom panel) were simulated. Each panel depicts all 100 curves of the ratios of $E_i/\hat{E}_i$ for all i. Vertical lines depict the 0.025th and 0.975th quantiles in each panel. The ratios are nearly identically unity in this range, implying negligible effects of CDF estimation.

where $\hat{F}_X$ denotes the estimated Gamma CDF. Density estimates for f_V are then obtained via the methods in Section 4.3). Our procedure is then applied with $r = 30$ and an $\alpha = 0.05$ level test (so that $T_\alpha = 28$ with associated Type I error $\alpha^* = 0.024$). The results are summarized in Table 3. False positives are sparse and the procedure again appears to work well.

5.4. *Simulation D.* With Type I error issues seemingly settled, we move to detection power. Our next simulation embeds two distinct clusters in IID uniform[0,1] random variables. Here, $\{V_i\}_{i=1}^n$ are IID Uniform[0,1] and $n = 4,000$. Next, we simulate two clusters — IID normal $\{C_j\}_{j=1}^{30}$ and $\{D_k\}_{k=1}^{30}$ — obeying

$$C_j \sim N(\mu = 0.25, \sigma^2 = 10^{-12}), \quad D_k \sim N(\mu = 0.75, \sigma^2 = 10^{-12})$$

The window length is set to $r = 30$ with $\alpha = 0.05$ one obtains $T_\alpha = 28$ and $\alpha^* = 0.024$. With the small variance chosen in the “alien" but normally distributed clusters, this scheme essentially injects spikes into the observation density at 0.25 and 0.75. Figure 5 shows a histogram of the 4,060 simulated data points with a bin width set to 0.05 (which is smaller than R’s default). Importantly, the two clusters are not visually obvious in this histogram. Table 3 reports additional specifications. Injecting clusters with larger variance will obviously

decrease the cluster detection rates; our point here is merely that we can detect clusters that are not visually apparent.

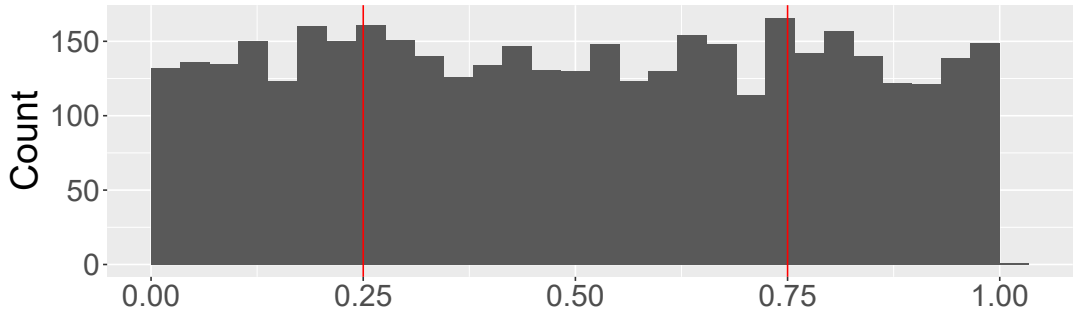


Fig 5: A histogram of the uniform density corrupted with alien clusters at 0.25 and 0.75. Density spikes are not visually evident.

Figure 6 shows our detected cluster locations and their frequencies over the 100 runs. Observe that both clusters are usually detected and that other “false positives" are sparse. Here, the methods have worked well.

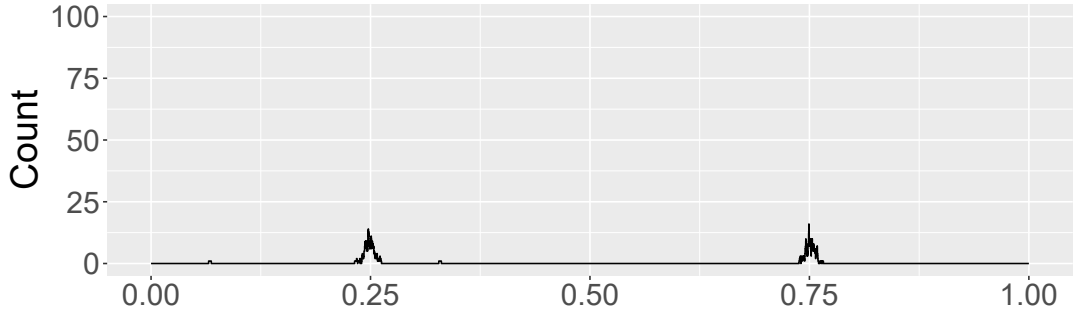


Fig 6: Detected clusters and locations when alien clusters at 0.25 and 0.75 are embedded into uniform[0,1] draws. The two clusters are detected at the correct location in all 100 runs.

5.5. *Simulation E.* Next, $n = 4,000$ independent draws of the density in (28) in Simulation C were generated. The sample is then corrupted with the above independent alien draws $\{C_j\}_{j=1}^{30}$ and $\{D_k\}_{k=1}^{30}$. Again, the window length is $r = 30$; for an $\alpha = 0.05$ level test, one obtains $T_\alpha = 28$ and $\alpha^* = 0.024$. Figure 7 shows the locations and numbers of the flagged clusters over 100 independent runs; again, performance is admirable. Note that while all clusters are detected, they are not precisely at the 0.25 and 0.75 quantiles since 0.25 and 0.75 are not the 25th and 75th quantiles of V with density given in Equation (28). This is not overly important; however, as the goal is to flag individual transactions within clusters, the procedure is successful in this regard. Table 3 provides additional specifications.

5.6. *Simulation F.* The next simulation demonstrates behavior when no adjustments for non-uniformity of f_V are made, describing a scenario where data is sampled from a density outside of the gamma family of distributions, but the method does not account for this departure. Our data here are simulated as in Simulation C; however, the density estimation methods

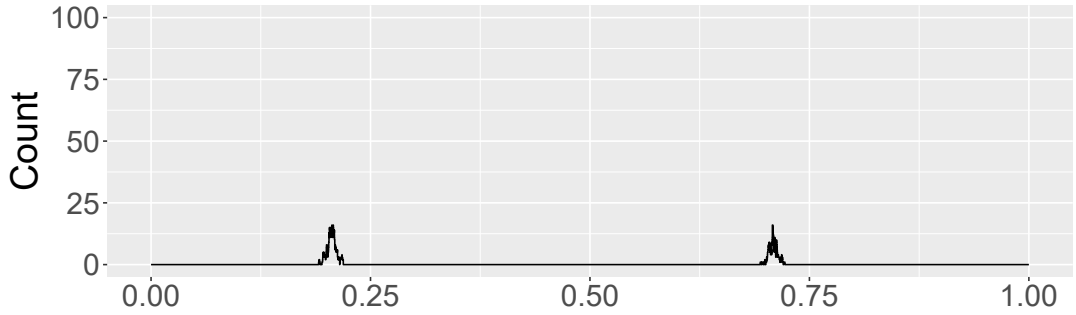


Fig 7: Detected clusters and their quantiles when alien clusters at 0.25 and 0.75 are embedded in draws from the density in (28).

Simulation Studies			
Simulation	False Positives	Detected Cluster 0.25	Detected Cluster 0.75
A	2	NA	NA
C	1	NA	NA
D	2	100	100
E	0	100	100
F	20	100	100

TABLE 3

Summary of false positives and detected clusters. Simulation B assessed independence, and clusters detection was not considered.

in Section 4.3 are eschewed. When the expected gap size in (14) does not account for non-uniform[0,1] data, many spurious clusters may arise from distributional mis-specification. Indeed, when $f_V(v) > 1$ for some $v \in (0,1)$, there are likely to be more observations near v and the expected gap size should decrease near such v. Failing to reduce the expected gap size will frequently induce $Y_i(\theta) = 1$ near such v and spurious clusters will be signaled. This dynamic is illustrated in Figure 8. Here, $n = 4,000$ IID samples from the density in (28) are sampled. Alien clusters $\{C_j\}_{j=1}^{30}$ and $\{D_k\}_{k=1}^{30}$ are then added to the sample. The window length is $r = 30$ for $\alpha = 0.05$ (again, $T_\alpha = 28$ and $\alpha^*0.024$). Note that many spurious clusters are detected at v with $f_V(v) > 1$. For this reason, one should not assume that $f_V(v)$ is uniform[0,1], or equivalently, that the Gamma fit is truth. This also demonstrates the veracity of our density estimation methods.

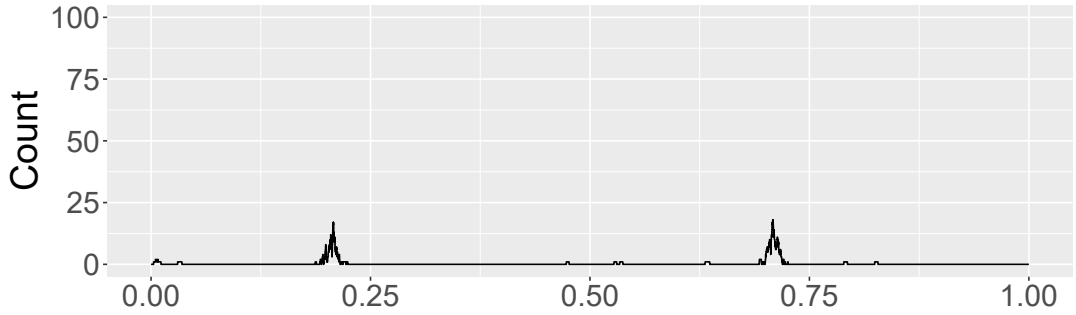


Fig 8: When the data is not uniform[0,1], the procedure detects many spurious clusters.

5.7. *Window length effects.* We now study window length effects. Here, 4,000 IID draws from the density in (28) were generated and corrupted with IID draws from $\{C_j\}_{j=1}^{30}$ and $\{D_k\}_{k=1}^{30}$, which essentially injects spikes into the density at 0.25 and 0.75. For each of the 100 independent data sets, we compute the number of times the scan statistic flags a cluster at 0.25 for various values of r (we omit study of the cluster at 0.75 as it is analogous). Figure 9 plots these counts over the 100 independent runs for various values of r. While good performance is seen from r values around 30, which is the true cluster size and is optimal in this sense, the methods also work well for a variety of distinct rs — from 15 to 50. In this case, one need not have to worry too deeply about selection of r.

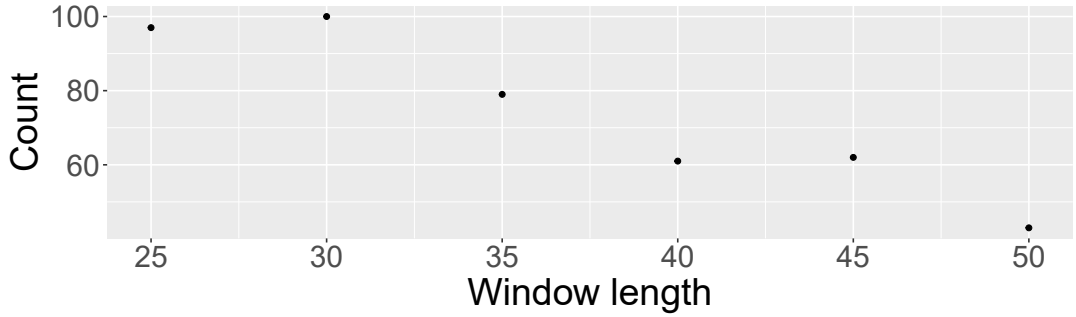


Fig 9: The number of times the cluster at 0.25 is detected.

6. Results. This section returns to our case studies of individual vendors and summarizes state-wide results for Mississippi.

6.1. *Case Studies.* Case studies of three vendors during a single FY are now supplied. The procedure constructed in Section 4 is applied to each vendor. Results on detected clusters and their rankings are reported. After applying the procedures in Section 4, close inspection of the transactions show that shoppers have a preference for a small number of popular items, examples could include pairing a sandwich and drink. These behaviors induce many clusters at small price points. Detected price point clusters depicting such preferences are hence disregarded when reporting results. We therefore ignore any detected cluster below $50.00. Any such cluster is not flagged or used in the ranking methods. This paradigm was maintained for all vendors.

First, we return to Vendor A, whose transactions were depicted in Panels (a) of Figures 2 and 3. This vendor conducted n =131,175 EBT transactions totaling $5,434,398.30. The Gamma fit produced parameter estimates of $\hat{\alpha} = 0.8239$ and $\hat{\beta} = 0.0199$. The interpolated density estimate of the V_is is plotted against a histogram in Figure 10. The partition in (16) is used for all vendors. The density estimate appears to match the histogram, except near the tail at unity. This is not an issue given the discussion in Section 4.5 (Simulation B) since the study is limited to transactions less than the 97.5th quantile. This also allows $n - i$ to be large enough for the convergence in Theorem 4.2 to apply.

The window length $r_d = 592$ was selected. The convention $\alpha = 0.05$ is adopted for all vendors. For Vendor A, the threshold $T_\alpha = 284$ with associated $\alpha^* = 0.043$ was used. The maximum value of the scan statistic never exceeded this threshold; hence, no clusters above $50.00 were detected. This vendor scores zero and the procedure is concluded.

The procedure was next applied to Vendor B, whose transactions were depicted in Panels (b) of Figures 2 and 3. Vendor B experienced $n = 12,987$ transactions totaling $519,523.67.

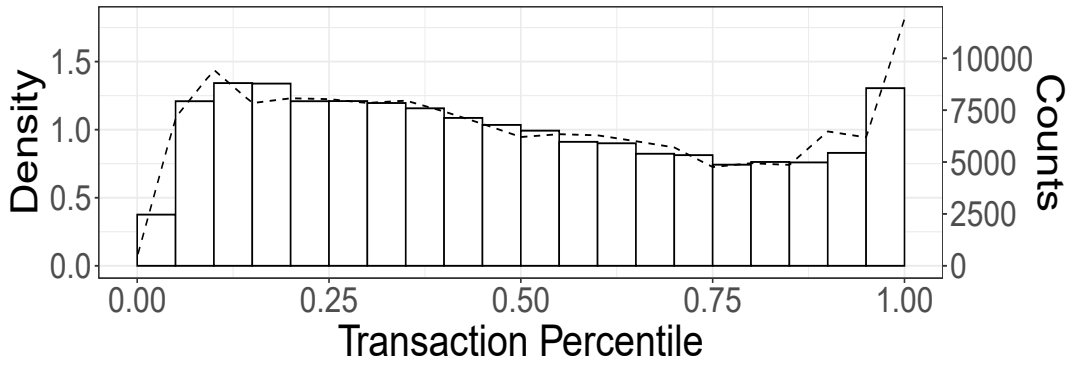


Fig 10: Histogram of the Gamma converted Vendor A transactions and its corresponding density estimate.

The initial Gamma fit produced parameter estimates of $\hat{\alpha} = 0.8640$ and $\hat{\beta} = 0.0216$. The CDF transformed density and histogram are plotted in Figure 11. The procedure selects a window length of $r_d = 58$ according to (22). For an $\alpha = 0.05$ level test, the threshold is $T_\alpha = 39$ with associated $\alpha^* = 0.040$.

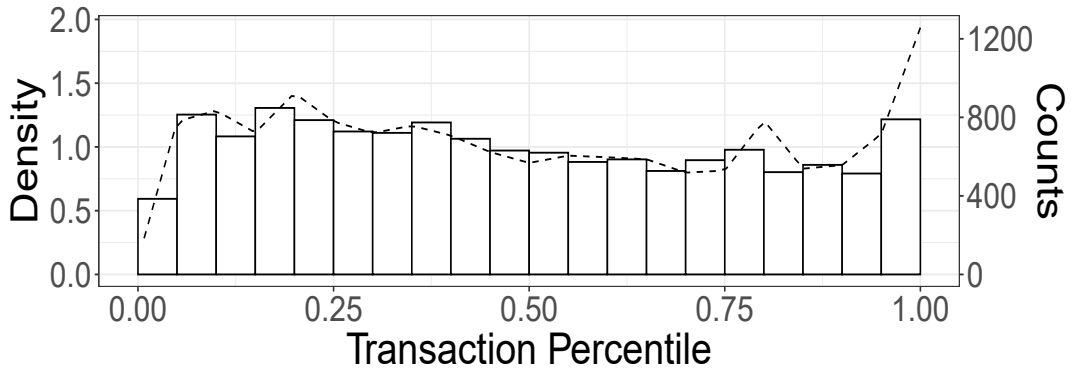


Fig 11: Histogram of the Gamma converted Vendor B transactions and its corresponding density estimate.

To aid intuition, the detected clusters are presented on the jittered transaction amounts $\{X_i\}$, not the CDF transformed data $\{V_i\}$. These are depicted in Figure 12, which shows a histogram with \$1 dollar bin widths. The transaction amounts appear on the x-axis and transactions counts are placed on the y-axis. Detected clusters appear in purple streaks based on the smallest θ where the scan statistic exceeds $T_\alpha = 39$. For a transaction amount X_i, lower values of θ having $S_n(r, i, \theta) > T_\alpha$ are shaded deeper red. The color-code key for θ is indicated in the vertical bar on the right. The largest θ depicted is $\theta_{\max} = 1/2$. One can visually see the width of the cluster as θ tends towards zero. Intuitively, the ranking method is the amount within the cluster multiplied by the area outlined by each colored spike. As previously discussed, our ranking method is based on the area in the spike to weight clusters with variance as more anomalous than those densely packed at a single dollar and cent amount.

Figure 12 shows clusters in integer multiples of \$65.00. Indeed, clusters are detected at the price points \$65, \$130, and \$195, as well as \$80 and \$100. The cluster locations, averages,

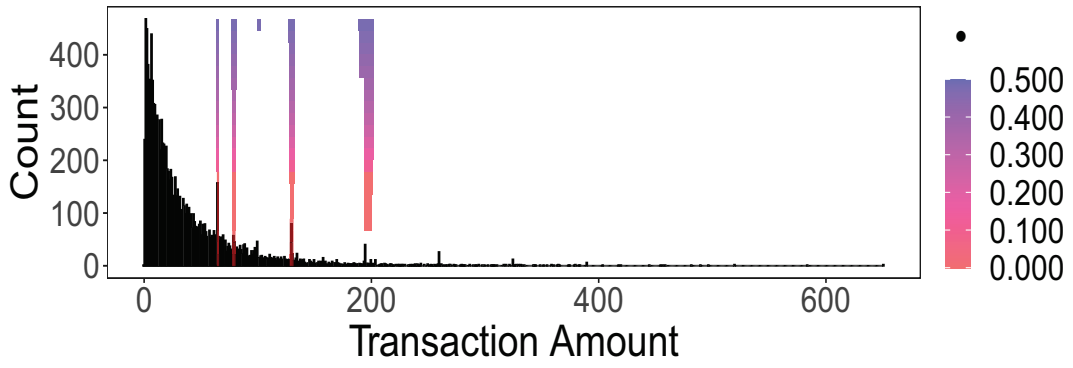


Fig 12: Several clusters are detected for Vendor B. The shaded regions are proportional to the cluster scores and are computed from (27).

and scores are summarized in Table 4. The procedure identified $m = 6$ clusters, and the vendor rank is $R_{(6)} = 55.81 \times 10^3$.

Vendor B				
Cluster i	τ_i	κ_i	Average Amount	Score: Thousands
1	\$189.00	\$201.45	\$195.43	55.81
2	\$127.68	\$132.27	\$130.03	26.54
3	\$77.64	\$81.07	\$79.37	10.70
4	\$64.51	\$65.47	\$65.00	4.95
5	\$99.92	\$102.00	\$100.50	0.30

TABLE 4

Summary of detected clusters and cluster ranks for Vendor B. Cluster scores are computed from (27).

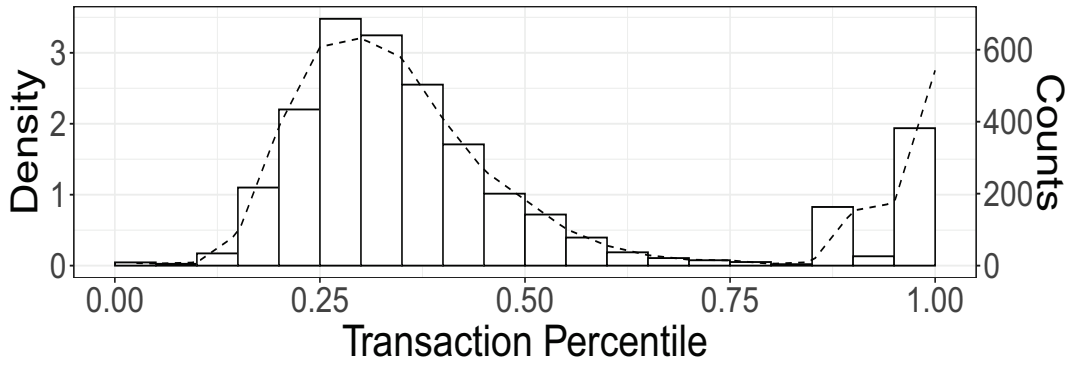


Fig 13: Histogram of the Gamma converted Vendor C transactions and its corresponding density estimate.

Next, we present results from a flagged vendor, called Vendor C, under the same format. Vendor C experienced 3,943 transactions totaling \$112,698.69, with the Gamma estimates

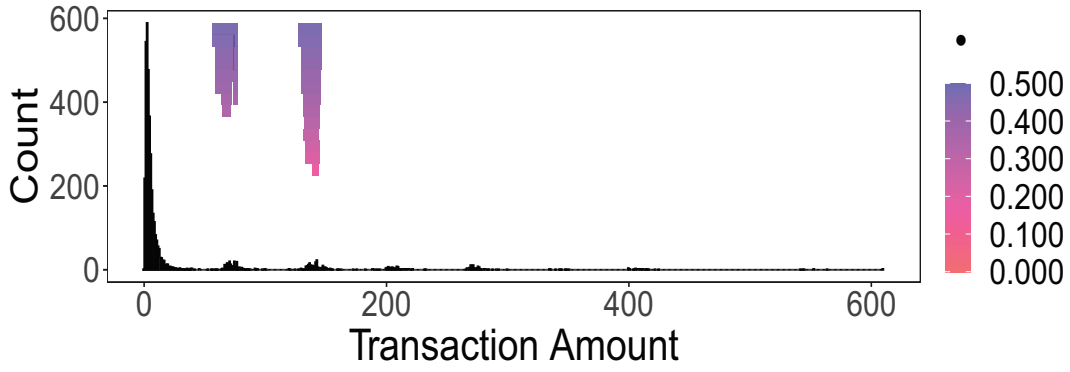


Fig 14: Vendor C clustering and ranking blocks. The shaded regions are proportional to the cluster scores.

Vendor C				
Cluster i	τ_i	κ_i	Average Amount	Score: Thousands
1	\$128.00	\$146.13	\$138.88	73.32
2	\$56.87	\$76.67	\$71.08	20.87

TABLE 5

Detected clusters for Vendor C. Scores are computed via (27).

$\hat{\alpha} = 0.43$ and $\hat{\beta} = 0.01$. A histogram of the Gamma CDF transformed transactions is displayed in 13. Vendor C is an outlier in terms of the Gamma fit; specifically, much of the density f_V between 0.60 and 0.80 is near zero. However, the procedure appears robust to the departure from the Gamma model and appears to work well. The class of linear densities in (15) seem particularly well suited for such extreme cases. The procedure chose $r_d = 44$. For an $\alpha = 0.05$ level test with $n = 3,943$ transactions, the threshold is $T_\alpha = 31$, which yields $\alpha^* = 0.026$. The cluster locations, averages, and scores are summarized in Table 5. The procedure identified $m = 2$ clusters, and the vendor score is $R_{(2)} = 73.32 \times 10^3$. Note that the ranking method weights these wider clusters more heavily than the tighter-packed clusters of Vendor B, as intended.

6.2. *Statewide Results.* The methods developed above were applied to all EBT accepting vendors from FY 2008-2017, where $n > 1,000$ transactions where experienced. This n facilitates the asymptotic independence of $\{\hat{Y}_i(\theta)\}$ in Subsection 5.2 (almost all vendors experienced more than $1,000$ EBT transactions). The reported amounts are the total excess amounts due to clustering. To compute these, suppose that the procedure detects m clusters at a specified vendor. Let A_j denote the total excess amount due to cluster j:

$$A_j = \left(\sum_{i=\tau_j}^{\kappa_j + r - 2} X_{(i)} \right) - n \int_{X_{(\tau_j)}}^{X_{(\kappa_j + r - 2)}} t \hat{f}_X(t) dt, \tag{29}$$

where $\hat{f}_X$ is the Gamma MLE fitted density. Then the total excess amount due to clustering for a vendor is $S = \sum_{j=1}^{m} A_j$ (or zero if no clusters are detected). The quantity S is computed for each vendor in the state and FY and summed statewide. Results are reported in Table 6, where "anomalous amount" is the excess due to clustering. To be cautious, the maximum θ

considered was 1/2. Blackstone's ratio — "better that ten guilty persons escape, than that one innocent suffer" — guided this stance.

Using the above parameterization, 10.47% of vendors were flagged. Table 6 depicts the total amount in dollars and as a percentage associated with clusters on a yearly basis.

Year	Anomalous Amount	Total Transaction Amount	Percentage
2009	$2,974,011	$645,011,310	0.46%
2010	$4,777,470	$839,013,298	0.57%
2011	$5,169,269	$908,550,428	0.57%
2012	$6,838,981	$959,879,314	0.71%
2013	$6,794,914	$949,182,221	0.72%
2014	$6,039,293	$885,047,604	0.68%
2015	$5,844,478	$874,198,495	0.67%
2016	$4,519,124	$759,449,273	0.60%
Total	$42,957,541	$6,820,331,943	0.63%

TABLE 6

Statewide aggregates of anomalous cluster amounts and statewide total transaction amounts are shown in the first two columns. Cluster amounts, as a percentage of total transaction amounts, populate the rightmost column.

From this information, anomalous clusters are detected, but only contain a small portion of the transactions. We are in agreement with the USDA that the program is an efficient mechanism to distribute benefits and likely has a misuse rate of less than 2%. Furthermore, systematic misuse is likely easy to identify. The top 100 cluster ranks were plotted in the histogram in Figure 15, which appears to approximate an exponential distribution.

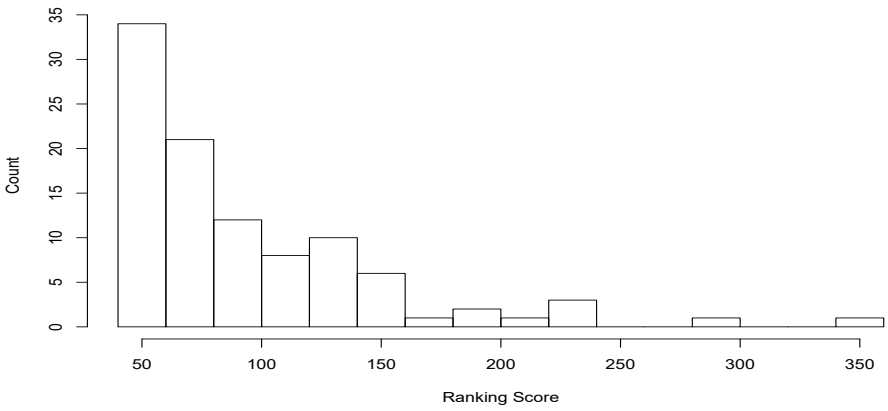


Fig 15: Histogram of ranking scores in units of thousands.

7. Conclusions and Comments. This study presented a forensic analysis of the state of Mississippi's SNAP program. Exploratory data analysis, news reports, and conversations with state investigators suggest fraudulent transactions that cluster near various pre-identified price points are of interest. Order statistic methods were developed and used to detect narrow clusters in the probability densities of the transactions. Our key asymptotic approximation to the gap sizes in the order statistics is given in (14). The methods used scan statistics to find

"local deviations" in a sequence of indicator variables calculated from the gaps of the order statistics. A method that ranks the severity of found clusters, based on their data depths, was also created.

Future directions for this work are numerous. Some relate to improving the ad-hoc choices used within. For example, changepoint methods may help set the partition in (16), subject to an increased computing expense. Other density estimates may also prove useful; however, the methods adopted here are robust and are adequate enough to work statewide. Also, methods that optimally select $\theta_{\max}$ seem worthy of investigation. Such methods would undoubtedly need regularity assumptions on f_V and are perhaps best handled on a case-by-case basis. Additionally, one may wish to develop peaks over threshold based methods to analyze extreme transactions. The methods could also be extended to accommodate multiple scanning windows, which is now common in scan statistic applications.

Deeper improvements include developing multivariate versions of our methods. Investigation of how to use the gap sizes, beyond say the simple zero-one indicators used here, are also important. Fraud structures other than density spikes may also arise. While the study here only looked for density spikes, modifying the methods to locate other types of fraud structures may prove useful. Also, locating density spikes may arise in non-fraud applications.

Some implications of this work are as follows. First, inexpensive and fast forensic statistical analysis can strengthen SNAP integrity, thereby increasing confidence that program funding is used as intended. Statisticians have the skills necessary to build robust and effective infrastructure for this purpose. Second, private vendors interested in selling SNAP integrity solutions are often overly critical of the integrity of the program (Dean, 2016); academic statisticians have the technical skills necessary to effectively evaluate proposed integrity solutions. Next, the vestige of the paper-based coupon books seems likely to be of economic and sociological interest. It is certainly interesting that clustering in the coupon book denominations continued a decade after the use of coupon books was discontinued. Finally, this analysis could be extended to the entire US, although differences in benefit disbursements among the states exist that could create differing price-point clusters or new patterns of interest.

The USDA estimates that approximately 1.5% of SNAP benefits are trafficked each year. From our work here, this quote seems like a reasonable estimate. In short, SNAP benefits appear to be overwhelmingly used as intended, contrary to some commonly held beliefs.

APPENDIX A: PROOF OF LEMMA 4.1

PROOF. First, we derive an expression for the CDF of G_i given $V_{(i-1)}$ in terms of the hazard functions. For each fixed i, observe that

$$P_{V_{(i-1)}}(G_i \le t) = 1 - P_{V_{(i-1)}}(\min(R_1, \ldots, R_{n-i+1}) > t) = 1 - [1 - P_{V_{(i-1)}}(R_1 \le t)]^{n-i+1}.$$

Applying (11) renders

$$P_{V_{(i-1)}}[G_i \le t] = 1 - \exp\{-(n-i+1)[H(V_{(i-1)}+t) - H(V_{(i-1)})]\}.$$

Letting $f_{G_i|V_{(i-1)}}(t)$ denote the corresponding probability density, differentiation yields

$$f_{G_i|V_{(i-1)}}(t) = \exp\{-(n-i+1)[H(V_{(i-1)}+t) - H(V_{(i-1)})]\}h(V_{(i-1)}+t)(n-i+1), \tag{30}$$

for $0 \le t \le 1 - v$.

Now combine

$$E_{V_{(i-1)}}\left[\frac{1}{h(V_{(i)})}\right] = \int_0^{1-V_{(i-1)}} \frac{1}{h(V_{(i-1)}+t)} f_{G_i|V_{(i-1)}}(t)dt$$

with (30) to finish our work:

$$E_{V_{(i-1)}}\left[\frac{1}{h(V_{(i)})}\right] = (n-i+1)\int_0^{1-V_{(i-1)}} \exp\{-(n-i+1)[H(V_{(i-1)}+t) - H(V_{(i-1)})]\}dt$$
$$= (n-i+1)\int_0^{1-V_{(i-1)}} P_{V_{(i-1)}}[G_i > t]dt$$
$$= (n-i+1)E_{V_{(i-1)}}[G_i].$$

□

APPENDIX B: PROOF OF THEOREM 4.2

PROOF. We need to show that $\lim_{n\to\infty} P_{V_{(\hat{i}_n-1)}}(Y_{\hat{i}_n}(\theta)=1) = 1-e^{-\theta}$ when $n\to\infty$ and $n-\hat{i}_n \to \infty$.

For any constant $\theta > 0$, Lemma (4.1) provides

$$P_{V_{(\hat{i}_n-1)}}(G_{\hat{i}_n} \le \theta E_{V_{(\hat{i}_n-1)}}[G_{\hat{i}_n}]) = 1 - P_{V_{(\hat{i}_n-1)}}\left(G_{\hat{i}_n} > \frac{\theta E_{V_{(\hat{i}_n-1)}}[h(V_{(\hat{i}_n)})^{-1}]}{n-\hat{i}_n+1}\right).$$

But since $P_{V_{(\hat{i}_n-1)}}(G_{\hat{i}_n} > c) = (1-F_{R_1|V_{(\hat{i}_n-1)}}(c))^{n-\hat{i}_n+1}$ and

$$P_{V_{(\hat{i}_n-1)}}[R_1 \le x] = [F_V(x) - F_V(V_{(\hat{i}_n-1)})]/[1-F_V(V_{(\hat{i}_n-1)})],$$

we obtain

$$P_{V_{(\hat{i}_n-1)}}(G_{\hat{i}_n} \le \theta E_{V_{(\hat{i}_n-1)}}[G_{\hat{i}_n}]) = 1 - \left[1 - \frac{\int_{V_{(\hat{i}_n-1)}}^{c} f_V(t)dt}{1-F_V(V_{(\hat{i}_n-1)})}\right]^{n-\hat{i}_n+1}, \tag{31}$$

where the upper integration limit is

$$c = V_{(\hat{i}_n-1)} + \theta\left(\frac{E_{V_{(\hat{i}_n-1)}}[h(V_{(\hat{i}_n)})^{-1}]}{(n-\hat{i}_n+1)}\right).$$

Since f is Lipschitz continuous and $f(v) > c > 0$ for all v, $1/h$ is also Lipschitz continuous over $(0,1)$. Using this and that $V_{(\hat{i}_n)} - V_{(\hat{i}_n-1)} \downarrow 0$ almost surely as $n\to\infty$ for $\hat{i}_n$, we obtain

$$E_{V_{(\hat{i}_n-1)}}[h(V_{(\hat{i}_n)})^{-1}] = h(v_{(\hat{i}_n-1)})^{-1} + o_p(1) \tag{32}$$

as $n\to\infty$. By Lipschitz continuity and $f(v) > c$, The same conclusion holds at each quantile $q(p)$ with $\hat{i}_n$ varying in n as long as $n-\hat{i}_n \to \infty$ as $n\to\infty$.

It now follows that

$$P_{V_{(\hat{i}_n-1)}}(G_{\hat{i}_n} \le \theta E_{V_{(\hat{i}_n-1)}}[G_{\hat{i}_n}])$$
$$= 1 - \left[1 - \theta\frac{E_{V_{(\hat{i}_n-1)}}[h(V_{(\hat{i}_n)})^{-1}]}{(n-\hat{i}_n+1)}\frac{f_V(V_{(\hat{i}_n-1)})}{1-F_V(V_{(\hat{i}_n-1)})} + o_p((n-\hat{i}_n+1)^{-1})\right]^{n-\hat{i}_n+1}. \tag{33}$$

Hence, (33) becomes

$$P_{V_{(\hat{i}_n-1)}}(G_{\hat{i}_n} \le \theta E_{V_{(\hat{i}_n-1)}}[G_{\hat{i}_n}]) = 1 - \left[1 - \frac{\theta}{n-\hat{i}_n+1} + o_p((n-\hat{i}_n+1)^{-1})\right]^{n-\hat{i}_n+1}$$
$$\to 1 - e^{-\theta},$$

where the convergence as $n - \hat{i}_n + 1 \to \infty$ may be found in Stirzaker (2003) among other texts. Since this result holds for all $q_p \in (0,1)$, all indicator functions in Equation (6) converge when $n - \hat{i}_n \to \infty$. □

APPENDIX C: DERIVATION OF GAP APPROXIMATION IN (14)

Combining (32) with Lemma (4.1) gives

$$(n - \hat{i}_n + 1) E_{V_{(\hat{i}_n - 1)}}[G_{\hat{i}_n}] = E_{V_{(\hat{i}_n - 1)}}\left[\frac{1}{h(V_{(\hat{i}_n)})}\right]$$

$$= \left[\frac{1}{h(V_{(\hat{i}_n - 1)})}\right] + o_p(1) \tag{34}$$

since $V_{(\hat{i}_n - 1)}$ and $V_{(\hat{i}_n)}$ both converge to $q(p)$ almost surely as $n \to \infty$.

Next define the empirical CDF of $\{V_i\}_{i=1}^n$ as $\hat{F}_V^{emp}(v) = \left(\frac{1}{n}\right) \sum_{i=1}^n 1_{[V_i \le v]}$, where $\hat{F}_V^{emp}(V_{(\hat{i}_n - 1)}) = (\hat{i}_n - 1)/n$ for each $\hat{i}_n$. Since $f_V(v) > 0$ for all $v \in (0,1)$, the Law of Large Numbers implies that $\hat{F}_V^{emp}(V_{(\hat{i}_n - 1)}) \to p$ almost surely as $n \to \infty$. Since $1/h$ is Lipshitz continuous, we have

$$\begin{aligned}
(n - \hat{i}_n + 1) E_{V_{(\hat{i}_n - 1)}}[G_{\hat{i}_n}] &= \left[\frac{1 - F_V(V_{(\hat{i}_n - 1)})}{f_V(V_{(\hat{i}_n - 1)})}\right] + o_p(1) \\
&= \left[\frac{1 - \hat{F}_V^{emp}(V_{(\hat{i}_n - 1)})}{f_V(V_{(\hat{i}_n - 1)})}\right] + o_p(1) \\
&= \frac{\left(\frac{n - \hat{i}_n + 1}{n}\right)}{f_V(V_{(\hat{i}_n - 1)})} + o_p(1) \\
&= \frac{(n - \hat{i}_n + 1)}{(n+1) f_V(V_{(\hat{i}_n - 1)})} + o_p(1).
\end{aligned}$$

Thus the approximation in Equation (14) holds since

$$E_{V_{(\hat{i}_n(q(p)) - 1)}}[G_i] = \frac{1}{(n+1) f_V(V_{(\hat{i}_n(q(p)) - 1)})} + o_p\left(\frac{1}{n - \hat{i}_n(q(p))}\right).$$

Acknowledgments. The authors thank two anonymous referees and the Editors for constructive comments that improved this paper. We also acknowledge helpful contributions of Leigh Ellen Barefield and Sheida Riahi.

Funding. Robert Lund was partially supported by NSF Grant DMS-2113592. This project was funded by USDA grant SNAP-RIIT-2015.

REFERENCES

ABDALLAH, A., MAAROF, M. A. and ZAINAL, A. (2016). Fraud detection system: A survey. *J. Netw. Comput. Appl.* **68** 90–113.

AGGARWAL, C. C. (2014). *Data Classification: Algorithms and Applications*. CRC press.

AGGARWAL, C. C. (2015). Outlier analysis. In *Data Mining* 237–263. Springer.

AHMED, M., MAHMOOD, A. N. and ISLAM, M. R. (2016). A survey of anomaly detection techniques in financial domain. *Future Gener. Comput. Syst.* **55** 278–288.

AHSANULLAH, M., NEVZOROV, V. B. and SHAKIL, M. (2013). *An Introduction to Order Statistics*. Springer.

AL-HASHEDI, K. G. and MAGALINGAM, P. (2021). Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. *Computer Science Review* **40** 100402. https://doi.org/10.1016/j.cosrev.2021.100402

ARNOLD, B. C., BALAKRISHNAN, N. and NAGARAJA, H. N. (2008). *A First Course in Order Statistics*. SIAM.

BOLTON, R. J. and HAND, D. J. (2002). Statistical fraud detection: A review. *Statist. Sci.* 235–249.

CANNING, P. and STACY, B. (2019). The Supplemental Nutrition Assistance Program (SNAP) and the economy: new estimates of the SNAP multiplier. Technical Report.

CHANDOLA, V., BANERJEE, A. and KUMAR, V. (2009). Anomaly detection: A survey. *ACM Comput. Surv.* **41** 1–58.

CLINE, D. R. AND AUSSENBERG R. A. (2018). *Errors and Fraud in the Supplemental Nutrition Assistance Program (SNAP)*. (CRS Report No. R45147). https://sgp.fas.org/crs/misc/R45147.pdf

COUNCIL, N. R. et al. (2013). *Supplemental Nutrition Assistance Program: Examining The Evidence to Define Benefit Adequacy*. National Academies Press.

DAVID, H. A. and NAGARAJA, H. N. (2004). *Order Statistics*. John Wiley & Sons.

DEAN, S. (2016). *SNAP: Combating Fraud and Improving Program Integrity Without Weakening Success: Hearings Before the Subcommittees on Government Operations and the Interior of the Committee on Oversight and Goverment Reform U.S. House of Represntatives*. 114th Congress. https://oversight.house.gov/wp-content/uploads/2016/06/2016-06-09-Stacy-Dean-Testimony-CBPP.pdf

DURTSCHI, C., HILLISON, W. and PACINI, C. (2004). The effective use of Benford's law to assist in detecting fraud in accounting data. *Journal of forensic accounting* **5** 17–34.

EKIN, T., IEVA, F., RUGGERI, F. and SOYER, R. (2018). Statistical Medical Fraud Assessment: Exposition to an Emerging Field. *International Statistical Review* **86** 379-402. https://doi.org/10.1111/insr.12269

FAULK, K. (2016). Alabama grocier to forfiet $5.2 million in food stamp fraud case. *Birmiham Real-Time News*.

FEIGIN, P. D. (1979). On the characterization of point processes with the order statistic property. *J. Appl. Probab.* **16** 297–304.

FELLER, W. (1966). *An Introduction to Probability Theory and Its Applications, Vol. 2*, first ed. Wiley, New York, NY.

FU, J. C. (2001). Distribution of the scan statistic for a sequence of bistate trials. *J. Appl. Probab.* **38** 908–916. MR1876548 (2002k:60040)

GOOD, I. and GASKINS, R. (1980). Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. *J. Amer. Statist. Assoc.* **75** 42–56.

HAIMAN, G. (2007). Estimating the distribution of one-dimensional discrete scan statistics viewed as extremes of 1-dependent stationary sequences. *J. Statist. Plann. Inference* **137** 821–828. https://doi.org/10.1016/j.jspi.2006.06.010 MR2301718

HAND, D. J., BLUNT, G., KELLY, M. G., ADAMS, N. M. et al. (2000). Data mining for fun and profit. *Statist. Sci.* **15** 111–131.

HE, Z., XU, X., HUANG, Z. J. and DENG, S. (2005). FP-outlier: Frequent pattern based outlier detection. *Comput. Sci. Inf. Syst.* **2** 103–118.

HILAS, C. S. and MASTOROCOSTAS, P. A. (2008). An application of supervised and unsupervised learning approaches to telecommunications fraud detection. *Knowl.-Based Syst.* **21** 721–726.

KARLIN, S. and TAYLOR, H. E. (1981). *A Second Course in Stochastic Processes*. Elsevier.

KONIJN, R. M., DUIVESTEIJN, W., KOWALCZYK, W. and KNOBBE, A. (2013). Discovering Local Subgroups, with an Application to Fraud Detection. In *Advances in Knowledge Discovery and Data Mining* (J. PEI, V. S. TSENG, L. CAO, H. MOTODA and G. XU, eds.) 1–12. Springer Berlin Heidelberg, Berlin, Heidelberg.

KOU, Y., LU, C.-T., SIRWONGWATTANA, S. and HUANG, Y.-P. (2004). Survey of fraud detection techniques. In *IEEE International Conference on Networking, Sensing and Control, 2004* **2** 749–754. IEEE.

LIAO, H.-J., LIN, C.-H. R., LIN, Y.-C. and TUNG, K.-Y. (2013). Intrusion detection system: A comprehensive review. *J. Netw. Comput. Appl.* **36** 16–24.

LIBERMAN, U. (1985). An order statistic characterization of the Poisson renewal process. *J. Appl. Probab.* 717–722.

LIU, X. and ZHANG, P. (2010). A Scan Statistics Based Suspicious Transactions Detection Model for Anti-money Laundering (AML) in Financial Institutions. In *2010 International Conference on Multimedia Communications* 210-213. https://doi.org/10.1109/MEDIACOM.2010.37

MAES, S., TUYLS, K., VANSCHOENWINKEL, B. and MANDERICK, B. (2002). Credit card fraud detection using Bayesian and neural networks. In *Proceedings of the 1st international naiso congress on neuro fuzzy technologies* 261–270.

MILLER, S. J. (2015). *Benford's Law*. Princeton University Press.
NAGLER, T. (2018). Asymptotic analysis of the jittering kernel density estimator. *Math. Meth. Stat.* **27** 32-46.
NAUS, J. I. (1982). Approximations for distributions of scan statistics. *J. Amer. Statist. Assoc.* **77** 177–183. MR648042
NGAI, E. W., HU, Y., WONG, Y. H., CHEN, Y. and SUN, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decis. Support Syst.* **50** 559–569.
PHUA, C., LEE, V., SMITH-MILES, K. and GAYLER, R. (2010). A Comprehensive Survey of Data Mining-based Fraud Detection Research. *CoRR* **abs/1009.6119**.
RAJ, S. B. E. and PORTIA, A. A. (2011). Analysis on credit card fraud detection methods. In *2011 International Conference on Computer, Communication and Electrical Technology (ICCCET)* 152–156. IEEE.
SCOTT, D. W. (1979). On optimal and data-based histograms. *Biometrika* **66** 605–610.
SHAO, M., LI, J., CHANG, Y., ZHAO, J. and CHEN, X. (2021). MASA: An efficient framework for anomaly detection in multi-attributed networks. *Computers & Security* **102** 102085. https://doi.org/10.1016/j.cose.2020.102085
SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London.
STIRZAKER, D. (2003). *Elementary Probability*. Cambridge University Press.
THIPRUNGSRI, S. and VASARHELYI, M. A. (2011). Cluster Analysis for Anomaly Detection in Accounting Data: An Audit Approach. *Int. J. Digit. Account. Res.* **11** 69–84.
TORGO, L. (2011). *Data Mining with R: Learning with Case Studies*. Chapman and Hall/CRC.
WAND, M. (1997). Data-based choice of histogram bin width. *The American Statistician* **51** 59–64.
WATSON, G. and LEADBETTER, M. (1964). Hazard analysis. I. *Biometrika* **51** 175–184.
WILSON, H. (2017). The extent of trafficking in the supplemental nutrition assistance program: 2012-2014. *Nutrition Assistance Program Report*.
WU, Q. and GLAZ, J. (2019). Robust scan statistics for detecting a local change in population mean for normal data. *Methodol. Comput. Appl. Probab.* **21** 295–314. https://doi.org/10.1007/s11009-018-9668-6 MR3915443