JOURNAL OF INFORMATION SYSTEMS
Vol. XX, No. XX
MONTH YEAR
pp. 1–20

American Accounting Association
DOI: 10.2308/ISYS-2023-066

External Nonfinancial Measures in Substantive Analytical Procedures: Contributions of Weather Information

Kyunghee Yoon
Clark University

Alexander Kogan
Miklos A. Vasarhelyi
Rutgers, The State University of New Jersey

Tim Pearce
KPMG UK

ABSTRACT: Recent studies and new standards suggest that auditors can use information from expanded external sources to verify their clients' financial information. We propose advanced substantive analytical procedures with external nonfinancial measures derived from weather information to test whether it helps detect misstatements. Using computational simulations and daily store-level sales data, we test whether the proposed procedures with weather indicators outperform the procedures without such indicators in substantive analytical procedures to identify overstated daily store sales. For the multilocation retail firm examined in this study, we find that the models with one or more weather indicators perform better at detecting misstatements than the models without them. When the reliability of relevant internal information is in question, the usefulness of weather indicators is apparent. Overall, our results provide evidence suggesting the potential value of external nonfinancial measures in auditing.

Keywords: external evidence; substantive analytical procedure; revenue testing; Big Data; exogenous data.

I. INTRODUCTION

Substantive analytical procedures (SAPs) are defined as those that assess the risk of material misstatement for account balances or classes of transactions based on analyses of plausible relationships among both financial and nonfinancial data (AICPA 2012). Auditing standards recommend using SAPs (AICPA 2012), and extant literature indicates that such procedures can uncover misstatements and irregularities, thereby improving the effectiveness of audit (Knechel 1988; Trompeter and Wright 2010). Nevertheless, the Public Company Accounting Oversight Board (PCAOB) (PCAOB 2014) has raised concerns about the poor performance of SAPs, especially for large income statement accounts such as revenues. Meanwhile, prior studies and audit practitioners point out fundamental difficulties in

We thank a large accounting firm for access to financial data and consistent support for this project, and Weather Underground (https://www.wunderground.com/) for access to weather data. This research has benefited from advice regarding weather information from Tony Broccoli and David Robinson and helpful comments from Theodore Mock, Carolyn B. Levine, and participants of the 2015 Strategic and Emerging Technologies Research Workshop and the 2016 Joint Midyear Meeting of the Accounting Information Systems Section and the Strategic and Emerging Technologies Section, especially Michael Alles (discussant). We are indebted to Ann Medinets, Steven Katz, Dimitri Lazo, and J. P. Krahel for their help in improving this manuscript.

The views expressed in this article are those of the author and do not reflect the official position of KPMG UK.

Kyunghee Yoon, Clark University, School of Management, Worcester, MA, USA; Alexander Kogan and Miklos A. Vasarhelyi, Rutgers, The State University of New Jersey, Rutgers Business School, Accounting Information Systems, Newark, NJ, USA; Tim Pearce, KPMG UK, Birmingham, United Kingdom.

Editor's note: Accepted by Senior Editor Tawei Wang.

Submitted: October 2023
Accepted: January 2024
Early Access: March 2024

developing accurate predictions for these accounts within performance materiality (Christensen, Elder, and Glover 2015; Glover, Prawitt, and Drake 2015; Yoon and Pearce 2021).

In July 2020, the AICPA's Auditing Standards Board issued *SAS 142: Audit Evidence* (AICPA 2020), recognizing the value of technology and information for modern audit procedures. The new standard highlights the expanding use of external information sources as audit evidence.

Recent accounting studies identify novel external information sources that auditors might use as audit evidence, such as weather records, social media, online news postings, and consumer demographic data, and note that they can offer incremental insights that more traditional audit evidence might not deliver (Yoon, Hoogduin, and Zhang 2015; Appelbaum, Kogan, and Vasarhelyi 2017). Also, some prior studies demonstrate that auditors need to incorporate audit evidence from various sources, such as external evidence that might not be controlled by managers and could offer economic and business insights (Maletta and Wright 1996; Trotman and Wright 2012). Although the new standard is designed based on the assumption that auditors should evaluate information to be used as audit evidence, including its authenticity and susceptibility to bias (Tysiac 2020), a limited number of studies demonstrate how and when auditors actually use external information sources.

By following the design science approach (Peffers, Tuunanen, Rothenberger, and Chatterjee 2007), we propose the design of SAPs utilizing external nonfinancial measures (ENFMs) and evaluate them by examining their use in detecting misstatements. In particular, the purpose of our proposed design is to demonstrate how to use weather indicators as an example of ENFMs to identify misstatements in a retailer's daily store sales and when auditors find their usefulness. Weather information's potential relevance can be found in certain industries based on the findings of existing literature (Li, Luo, Zhang, and Wang 2017; Badorf and Hoberg 2020). The provenance of this information from the National Oceanic and Atmospheric Administration (NOAA), an unbiased third party, enhances its reliability.

Although prior studies find the associations between weather conditions and sales of specific industries, these findings do not assure that they have predictive powers for audit purposes. Sales are influenced by various factors other than weather conditions including market competition, a firm's reputation and marketing strategies, and others. Auditors, who can access a client's internal information, might not find any incremental value of weather information. It is uncertain how to use weather information to predict sales. Weather conditions are measured in various ways. For example, temperature can be measured by air temperature, apparent temperature, heating degree days, and others. Or the mean or changed temperature might be relevant to sales. Also, relevant weather measures can be different depending on the firm, and weather measures are relevant in certain cases only, such as in specific regions.

Consistent with prior studies proposing advanced SAPs (Allen, Beasley, and Branson 1999; Kogan, Alles, Vasarhelyi, and Wu 2014), we propose SAPs for disaggregated financial data. Although external auditors are not obligated to assure daily store level sales, PCAOB inspectors have expressed concerns regarding auditors' failure to examine location-specific revenue accounts that had specific risks (PCAOB 2014), and expectations derived from a detailed level are likely to enhance the probability of detecting misstatements (AICPA 2012). Additionally, the proposed approach can offer valuable insights to not only external but also internal auditors interested in incorporating external evidence to examine disaggregated financial data. We use daily store sales from a large, publicly held, multilocation U.S. retailer, and weather information from Weather Underground. We mainly employ apparent temperature, measured by the joint effects of air temperature, relative humidity, and wind speed, as the primary weather indicator and also adopt heating and cooling degree days and precipitation, which are used by the National Weather Service (NWS) and prior studies (Starr-McCluer 2000; Pérez-González and Yun 2013). Seasonal autoregressive (AR) integrated moving average (MA) with exogenous factors (SARIMAX) models are adopted to forecast one-day-ahead store sales. The daily sales of a peer store from the target firm, with similar sales patterns from a prior year, are used as relevant internal financial information and included as a control variable in analytical models. Peer store sales are strongly correlated with target store sales as the target firm provides similar services and products in all of its stores. A firm's reputation and firm-wide marketing strategy are largely controlled by utilizing peer store sales (Allen et al. 1999).

We evaluate whether the proposed SAPs are superior by adopting computational simulations. We seed material overstated misstatements in actual one-day-ahead store sales and test whether the proposed SAPs with one or more weather indicators outperform the SAPs without them. Importantly, to modify the reliability of internal financial data, two different types of misstatements are examined: coordinated and uncoordinated misstatements. Coordinated misstatements are defined as incidents in which both target store sales and peer store sales are misstated (i.e., weak internal control). Uncoordinated misstatements are defined as incidents in which only the target store sales are misstated (i.e., moderate or strong internal control).

Our computational simulation results generally match our expectations. SAPs with peer sales and one or more weather indicators are superior to the SAPs with peer sales only. Weather-conscious SAPs marginally outperform the benchmark SAPs under uncoordinated misstatement assumptions. However, in a setting with coordinated misstatements, the usefulness of the weather information is especially apparent. We perform additional analysis with different research settings and modified SAPs. The results are generally consistent. We also study if the usefulness of weather records varies by region. The weather-informed SAPs are superior to the benchmark SAPs in all regions, especially the Northeast when apparent temperatures are considered.

The contributions of this study are three-fold. First, it shows the use of ENFMs in auditing and in a condition where the benefits of ENFMs are apparent. Existing literature tests the impact of different types and sources of audit evidence by studying nonfinancial information (Cohen, Krishnamoorthy, and Wright 2000; Brazel, Jones, and Prawitt 2014) and external information (Maletta and Wright 1996; Trotman and Wright 2012). However, very few ENFMs derived from external Big Data are studied. Several recent studies propose the potential for external information from Big Data in auditing (Appelbaum et al. 2017) conceptually, but only a few studies present its potential by studying data (e.g., Rozario, Vasarhelyi, and Wang 2023). We examine the performance of SAPs when both internal financial and external nonfinancial information are considered simultaneously. Our computational simulation results show the usefulness of ENFMs in auditing that are relevant to the target firm's performance. Also, we show that consistent with Wright and Ashton (1989), the extent of that usefulness in misstatement identification differs depending on the reliability of internal evidence.

Second, this study extends prior proposed approaches to enhance SAPs by incorporating ENFMs, especially those derived from weather records. Improving the effectiveness of SAPs, especially for large income statement accounts, such as revenues, is increasingly critical for audit practice (PCAOB 2013, 2014). However, there are inherent difficulties in predicting revenues within performance materiality (Glover et al. 2015; Yoon and Pearce 2021). Prior studies propose ways to improve the performance of analytical procedures by incorporating external information, such as macroeconomic (Lev 1980), industrial data (Hoitash, Kogan, and Vasarhelyi 2006), and social media data related to customer interest and satisfaction (Rozario et al. 2023). However, the ENFMs this study proposes, weather indicators, have benefits previously proposed measures may not offer. Unlike weather records, microeconomic and industrial data may not be appropriate for examining timely and locational high-frequency financial records, which can enhance SAPs (Allen et al. 1999) and become available after considerable time has passed. Measures from social media data have reliability and quality concerns, unlike weather records that are maintained by the NOAA.

Finally, this study contributes to prior research exploring the relevance of weather information for measuring firm performance (Pérez-González and Yun 2013; Kirk, D. Stice, and H. Stice 2022). For instance, Badorf and Hoberg (2020) find the meaningful explanatory and predictive power of weather information on the daily sales of a German retailer. We extend this literature by testing whether weather information can be useful in detecting misstated financial performance.

The next section discusses related issues in auditing practice, summarizes the relevant literature, and presents our research questions. Section III presents the research design, and Section IV describes the computational simulation results. The last section discusses the results and makes recommendations for future research.

II. BACKGROUND, MOTIVATION, AND RESEARCH QUESTIONS

Design Science Research Method

This research follows a design science research method (DSRM) (Peffers et al. 2007) by proposing novel SAPs that support the internal and external audit of financial statement accounts. The empirical research method, which is most frequently adopted by prior accounting literature, is based on measuring social phenomena. On the other hand, since computer science and engineering studies value applicable problem solutions, the DSRM is frequently employed by these domains (Peffers et al. 2007). Although the DSRM is not widely adopted by accounting literature, a growing number of recent accounting information systems studies, especially examining audit data analytics (Christ, Emett, Summers and Wood 2021) and continuous auditing (Yoon, Liu, Chiu, and Vasarhelyi 2021) employ this method. Kogan, Mayhew, and Vasarhelyi (2019) demonstrate that it is reasonable to utilize a similar approach to the approach engineering and computer science adopt for evaluating the usefulness of data analytics in the accounting domain because engineering and computer science are closely relevant in the development of data analytics.

Prior studies developing analytical procedures often employ the DSRM (Allen et al. 1999; Vandervelde, Chen, and Leitch 2008; Kogan et al. 2014) although those studies do not explicitly state they adopt the DSRM. They identify problems and motivations prior studies do not solve, propose advanced artifacts that enhance the effectiveness of analytical procedures, and evaluate their advantages by evaluating the artifact as studies utilizing the DSRM do. Those studies frequently evaluate their proposed SAPs using single or several firm data to constrain the setting (e.g., business process or industry) (Allen et al. 1999; Vandervelde et al. 2008) because it is difficult to generalize the superiority of specific SAPs, which need to be developed in a tailored way for each firm. Proposed procedures often require disaggregated financial data,[1] and a limited number of proprietary cases are employed (Allen et al. 1999).

Current Issues with SAPs

When performing analytical procedures, auditors develop expectations for a client's reported financial statement data, compare their expectations to the reported data, and investigate any significant differences (AICPA 2012). Preliminary analytical procedures are used to determine the nature, timing, and extent of other audit procedures, and review-stage analytical procedures are used to evaluate the collected evidence. Auditors use SAPs to detect material misstatements for specific accounts or classes of transactions.

However, PCAOB inspectors have expressed concern about auditors' poor performance in SAPs, particularly for significant income statement accounts, such as revenue (PCAOB 2013, 2014). Meanwhile, auditors often judge that it is inherently unlikely for SAPs to provide high assurance levels because the performance materiality for these accounts is very small (Glover et al. 2015). Although prior studies propose various approaches that incorporate advanced analytical models and financial and nonfinancial information for SAPs, Yoon and Pearce (2021) find that previously proposed approaches barely achieve high assurance for revenue.

ENFMs in Analytical Procedures

Advances in data and data analytics dramatically influence audit practice. Large accounting firms invest billions in technology (Kapoor 2020), and the CPA exam is transforming to reflect accountants' new prerequisite skills and knowledge, especially those related to data, automation, and data analytics (National Association of State Boards of Accountancy (NASBA) 2021). The AICPA has also announced new standards for audit evidence to address the expanding sources of information to be utilized as audit evidence (AICPA 2020). The old standard was developed based on considerations of paper documents mainly from clients' internal sources. Since technologies affect business operations, the new standard recognizes different forms of audit evidence, including ones from external sources.

Recent accounting studies address the influence of data and data analytics in auditing. Advanced information technologies allow auditors to access, store, and process large amounts of data, often from external sources, such as media, Internet of Things (IoT), locational, videos, audio, and weather, frequently called Big Data, as audit evidence (Appelbaum et al. 2017). Prior studies predict that evidence from these novel sources might have the potential to dramatically change the assurance function largely due to their public nature and expanding scope (Brown-Liburd and Vasarhelyi 2015; Yoon et al. 2015). Some recent studies test whether data analytics enhances audit effectiveness (Christ et al. 2021; Barr-Pulliam, Brown-Liburd, and Sanderson 2022), but few studies focus on the data itself (e.g., Rozario et al. 2023) and on especially circumstances on where the data can be used in auditing.

Using ENFMs in analytical procedures could enhance the recognition of errors or misstatements in financial statements for several reasons. ENFMs are not affected by internal accounting errors or irregularities. Because management can be motivated to manipulate their firm's account balances to align with industry trends and the firm's budget (Beneish 1997), it is possible to miss fraudulent behavior when using only internal information provided by management. Similarly, based on the PCAOB inspection reports, Griffith, Hammersley, and Kadous (2015) point out that auditors tend to rely too much on a client's data and sometimes fail to reconcile conflicting evidence.

ENFMs can broaden auditors' thinking because they provide unique insights into account balances that existing measures, especially internal financial measures, might not offer. ENFMs can broaden auditors' understanding of external factors associated with the economic and business environment that influence firm performance like customer satisfaction and behavior, economic conditions, or market competition. ENFMs might be particularly useful for large income statement accounts, such as sales and net income, which are significantly subject to external influences.

We are aware of very few existing studies examining the use of ENFMs derived from novel data in analytical procedures with relevant data, except for Rozario et al. (2023), which propose a way to incorporate Twitter information in SAPs for monthly firm sales of business-to-customer industries. They introduce Twitter-based measures of customer interest and satisfaction and find the usefulness of the indicator associated with customer interest in SAPs. In addition, a few prior studies propose the use in analytical procedures of external information from industrial and economic data (Lev 1980; Hoitash et al. 2006). For instance, Lev (1980) demonstrates that considering macroeconomic indicators like gross national product (GNP) and total corporate profits can help understand the environment in which a client operates. He shows that models with these economic indicators are superior to the benchmark model for sales, operating

[1] Since unlike preliminary and review-stage analytical procedures, SAPs are employed as audit evidence for a single account or a group of transactions, SAP might need to be conducted with disaggregated financial data such as general ledger, store or product line level data.

income, and net income. Hoitash et al. (2006) propose a novel approach to identify relevant industrial data from peer firms selected by using SIC codes and financial measures. They find that incorporating peer firms' contemporaneous financial information significantly enhances SAPs for sales and cost of goods sold.

However, measures proposed by those studies have limitations. The contemporaneous macroeconomic indicators used by Lev (1980) might not be available to auditors until considerable time has passed. For example, the U.S. GNP of Q4 2019 became publicly available on March 26, 2020 (Bureau of Economic Analysis (BEA) 2020). Similarly, peer firms' contemporaneous financial records might not be immediately available to auditors. Hoitash et al. (2006) focus on the situation where peer firms and a target firm share the same audit firm, allowing them to assume that auditors can share relevant information within their firm. Even if their assumption is valid, there is no guarantee that audited financial information for peer firms is ready for the target firm's auditor. By contrast, ENFMs derived from internet sources are often updated frequently. Since weather records, social media, and locational data are updated promptly, auditors can access them to examine transactions and accounts in almost real-time. Also, macroeconomic indicators and peer firm data proposed by previous studies might not be suitable for analytical procedures for disaggregated financial data on weekly, store, or regional sales, although examining disaggregated data allows auditors to effectively identify misstatements or errors in financial records (Allen et al. 1999). Finally, the measures derived from the social media data proposed by Rozario et al. (2023) have reliability concerns because the quality of data is not validated and is generated by specific individuals who might not represent the entire population of customers.

Weather Records as an ENFM in Analytical Procedures

Are Weather Records Relevant to Firm Performance?

We expect that weather information, as an example of ENFMs, can be relevant evidence for specific industries. To use weather records in analytical procedures, auditors need to evaluate whether they provide information that underlies firm operations and performance. Weather information could be relevant audit evidence to examine sales of certain industries like retail, utilities, and transformation based on the findings of prior studies (Starr-McCluer 2000; Forbes and Lederman 2010; Pérez-González and Yun 2013). For instance, weather might affect the sales of retailers for the following reasons: it can affect the customers' needs for products (Štulec, Petljak, and Naletina 2019), have psychological effects on customers (e.g., the influence of promotions) (Li et al. 2017), or impact customers' visits to stores (Starr-McCluer 2000). Starr-McCluer (2000) expects that unfavorable weather conditions might interrupt customers' store visits, thereby postponing sales, and finds that unusual weather conditions explain the monthly sales fluctuations of retailers, notably those selling durable goods. Anecdotal evidence concerning the economic influence of weather on business activities can be found in financial reports. Management often discusses risks related to weather in management discussion and analysis (MD&A), offering an overview of the year's operations, and in Item 1A: Risk Factors, containing information about the most significant risks that influence a firm's business. The weather is often blamed for an organization's poor performance. For instance, Peak Resorts, Inc. (2015) states in its 2015 10-K report, "Weather events…impacted two of the three major holiday periods of the 2014/2015 ski season and adversely affected the ski industry in general."

Weather information can also be relevant to location- and time-specific financial performance, such as weekly state sales or daily store sales. Since weather conditions vary widely by time and location, weather information can be used to verify time- and location-specific firm performance. Although the associations between time- and location-specific firm performance and weather are complicated (Badorf and Hoberg 2020) largely due to fluctuations in both datasets, this might show its relevance closely since the impact of weather on firm performance can be offset when more aggregated data is examined (Starr-McCluer 2000).

Are Weather Records More Useful When Internal Evidence Is Less Reliable?

The reliability of audit evidence is influenced by its provenance. If the information is generated by a client, auditors must consider the potential risks of management bias. If the information is obtained from third parties or external sources, risks to its completeness and accuracy become more relevant (AICPA 2018). However, weather information is generated and maintained by the NOAA, a regulatory agency that conducts periodic quality control to ensure the accuracy of the information it provides (Hubbard, Guttman, You, and Chen 2007). We expect that weather records have low risks concerning their completeness and accuracy.

Some prior studies show that the usefulness of external evidence is affected by the level of reliability in internal evidence, collected from a client with moderate or strong internal control versus weak internal control (Wright and Ashton 1989; Maletta and Wright 1996). Wright and Ashton (1989) demonstrate that when a client's internal control is weak, external evidence tends to identify more errors than internal evidence. Similarly, Trotman and Wright (2012) find that

American Accounting Association

auditors are less likely to rely on external evidence in their fraud risk assessment when internal evidence shows low fraud risk. The usefulness of external evidence in SAPs might be more significant when internal evidence may have been tampered with than when it is not.

III. DESIGN OF SAPS UTILIZING WEATHER INFORMATION

SAPs are conducted as following steps: (1) developing an expectation, (2) deciding a tolerable difference, (3) comparing the difference between the expectation from step (1) and the reported amount, (4) investigating substantial difference, which is decided in step (2), and (5) revising the exception, concluding the SAP cannot provide the sufficient appropriate audit evidence, or additional procedures are conducted. The design of SAP we propose in this study mainly focuses on step (1).

Background

We propose SAPs with weather information by using a retail firm's store-level sales data from fiscal years 2011 and 2012, obtained from a global accounting firm. Consistent with prior studies proposing advanced SAPs, we use single firm data because the proposed SAPs require disaggregated store-level financial data only proprietarily available. We design SAPs for daily store-level data for several reasons. First, expectations derived from a detailed level often have a greater possibility of detecting misstatements than making broad comparisons (AICPA 2012), although it is challenging to create a stable model. Since the risk that material misstatement could be concealed by offsetting factors increases as a client's operations become more complex and diversified, examining disaggregated data helps to reduce such risk (AICPA 2012). The PCAOB (2014) has raised concerns regarding auditors' failure to examine location-specific revenue accounts, even though it could help to identify misstatements. However, prior analytical procedure studies frequently examine monthly firm-level financial data, mathematically inferred from quarterly firm-level data (Hoitash et al. 2006; Chen and Leitch 1999; Rozario et al. 2023), possibly due to data access difficulties. Second, the use of location- and time-specific weather information as audit evidence for store sales might provide useful insights for internal and external audit practitioners who are interested in external evidence suitable for disaggregated financial records, updated in near real-time. As auditors adopt advanced technologies, auditing on a continuous, real-time basis becomes more feasible. In response, *SAS 142: Audit Evidence*, describes that information about transactions, which is available on a continuous, real-time basis, can be used as audit evidence by developing procedures with automated tools (AICPA 2020).

The targeted firm is a publicly held, multilocation retailer that mainly sells durable goods and has homogeneous worldwide operations, although only observations from the continental United States are used in this study. The firm endeavors to offer consistent products and services across stores, although store sizes differ. The dataset[2] includes the store-specific daily sales and each store's address. Over the targeted fiscal years, the external auditor expressed an unqualified opinion on the firm's annual reports and did not express concern regarding the effectiveness of the target firm's internal controls. Even though daily balances were not audited separately, the auditors did not find signals of material misstatement in the sales revenue accounts.

Analytical Model Specification

Prior studies suggest sophisticated statistical methods, such as artificial neural networks (Koskivaara 2004), three-stage least squares (Leitch and Chen 2003), and automated equilibrium correction modeling (Omura and Willett 2006) for enhancing analytical procedures. Kogan et al. (2014) compare the broadest range of statistical models, such as simple linear regression, simultaneous equation modeling, vector AR modeling (VAR), and generalized AR conditional heteroskedasticity. They conclude that simple linear regression and VAR outperform other models. By examining the findings of prior studies examining analytical procedures for revenue accounts, Yoon and Pearce (2021) show the benefits of time series models in enhancing analytical procedures, although it is difficult to identify the best-performing models from the findings of prior studies that often examine unique datasets, and it is difficult to develop generalized analytical model applicable for a wide variety of firms. Accordingly, we do not adopt previously proposed models because frequently these models are not suitable for our research setting but use one with time series components.

[2] Due to our limited access to the proprietary data, we cannot consider other financial and nonfinancial information for analytical models, resulting in the limitation of this study. For instance, Hoitash et al. (2006) and Rozario et al. (2023) use accounts receivable to predict monthly firm-level sales. However, because accounts receivable is primarily driven by sales, we do not agree that it is a relevant control variable for sales predictions. Our control variable, peer store sales, is closely associated with target store sales, and we use time series models, as we will describe further. Consequently, we assume that our analytical models might be sufficient to evaluate the incremental values of weather information. However, we do not argue that our proposed models are the best models that auditors can develop, because auditors can obtain a wide variety of internal information.

The SARIMAX model is adopted in this study because some prior studies, forecasting daily fluctuations with exogenous variables, adopt this model (Cools, Moons, and Wets 2009; Arunraj, Ahrens, and Fernandes 2016). In the series $Y_{j,t}$, if the tth observation of the dependent variable, daily sales for store j is modeled, it can be represented by the following equation:

$$\phi_p B \Phi_P B^s (1-B)^d (1-B^s)^D Y_{j,t} = \beta X_t + \theta_q B \Theta_Q B^s e_t$$

where s is the length of seasonality; ϕ_p B represents the order of the nonseasonal AR component, and $\theta_q B$ indicates the nonseasonal MA component. $\Phi_P B^s$ and $\Theta_Q B^s$ indicate the seasonal AR and the seasonal MA polynomial, respectively. $(1-B)^d$ $(1-B^s)^D$ indicates nonseasonal and seasonal time series differing components. The parameters p, d, and q are the order of AR, the degree of difference, and the order of the MA, respectively. P, D, and Q are the corresponding parameters of the seasonal part of the model. e_t represents prediction error terms, and B is the backshift operator on Y_t. β is a vector with the regression coefficients corresponding to predictors, and X_t is the matrix of exogenous variables. Often, a SARIMAX model is represented as SARIMAX $(p, d, q)(P, D, Q)_S$.

We find a strong week-to-week effect based on the partial autocorrelation function. For instance, last Monday's sales and this Monday's sales have similar patterns. Hence, the model we eventually adopt is SARIMAX $(1, 0, 0)(1, 0, 0)_7$ with exogenous variables to predict the daily store-level sales revenue balance.[3] To test the research questions, we compare misstatement detection performances derived from SARIMAX $(1, 0, 0)(1, 0, 0)_7$ with *PEER*, called the benchmark model, and SARIMAX $(1, 0, 0)(1, 0, 0)_7$ with *PEER* and weather indicator(s), called the weather model hereafter.

Weather data for this study are obtained using the Application Programming Interface of historical weather information from Weather Underground. Its information comes from the NOAA's NWS and its personal weather stations, and it is used in prior studies (Li et al. 2017). Using the zip code of each store, we collect relevant daily climate data.

Weather Indicators

To examine weather conditions associated with daily store sales, we primarily adopt a weather variable associated with temperature perceived by humans, called Apparent Temperature (*AT*). *AT* can be determined by a combination of Heat Index (*HI*) and Wind Chill Index (*WCI*), as used by the NOAA's NWS. These indices are used to evaluate the influence of weather conditions on business performance by prior studies (e.g., Arunraj and Ahrens 2015). When the temperature at a particular grid point falls to 50°F or less, *WCI* is used to calculate *AT*. When the temperature rises above 80°F, *HI* is used. Between 51°F and 80°F, *AT* will be the ambient air temperature, so the average temperature is used for *AT*. *HI* and *WCI* reflect that air temperatures are modified by humidity and wind, respectively, as influencing the perception of temperature. For instance, higher humidity can contribute to a warmer perceived temperature. Similarly, existing literature demonstrates the impact not only of temperatures but also of humidity and wind speeds on retail sales (Badorf and Hoberg 2020; Rose and Dolega 2022). *HI* is the level of discomfort caused by the combined effects of air temperature and humidity (Rothfusz 1990), calculated as follows:

$$\begin{aligned} HI = & -42.379 + 2.04901523 \times T + 10.14333127 \times RH - 0.22475541 \times T \times RH \\ & - 0.00683783 \times T^2 - 0.05481717 \times RH^2 + 0.00122874 \times T^2 \times RH + 0.00085282 \times T \\ & \times RH^2 - 0.00000199 \times T^2 RH^2 \end{aligned}$$

where T is the air temperature in °F, and RH is the relative humidity as a percentage.[4] *WCI* is estimated by the air temperature and wind speed. It is useful in determining the danger from winter winds and freezing temperatures. *WCI* (Osczevski and Bluestein 2005) is calculated as follows:

$$WCI = 35.74 + 0.6125 \times T - 35.75 \times V^{0.16} + 0.4275 \times T \times V^{0.16}$$

where V is the wind velocity in mph, and T is the air temperature in °F.

As an alternative measure of *AT*, we examine heating degree days (*HDD*) and cooling degree days (*CDD*), which are often adopted in prior studies as well (Starr-McCluer 2000; Pérez-González and Yun 2013). When the average daily

[3] We adopt the model SARIMAX $(1, 0, 0)(1, 0, 0)_7$, largely because we can straightforwardly reflect a week-to-week effect in a model, but this is not the best performing model among the models we test. Using a subset of randomly selected stores, we examine various SARIMAX models and find several other models that marginally outperform SARIMAX $(1, 0, 0)(1, 0, 0)_7$, including SARIMAX $(1, 0, 1)(1, 0, 0)_7$ and SARIMAX $(1, 1, 1)(0, 1, 1)_7$. In Appendix C, we report the major empirical results derived from SARIMAX $(1, 0, 1)(1, 0, 0)_7$ and SARIMAX $(1, 1, 1)(0, 1, 1)_7$. The results are generally consistent with the results we find with the subset of stores, and importantly are consistent with the major findings in this study.

[4] The formula to calculate *HI* is modified depending on *RH* and temperature. The detailed formulas are described in Appendix B.

temperature exceeds 65°F, the difference between 65°F and the average daily temperature is considered as *CDD* since air conditioning might be used to reduce the temperature. Similarly, *HDD* is calculated as cooling degree days when the average daily temperature is below 65°F. The average daily temperature is measured by the maximum daily temperature plus the minimum daily temperature divided by two.

In additional analyses, we also consider other weather variables related to unfavorable weather conditions. *PRECIP* is a dummy variable that is coded as 1 if the level of precipitation is greater than zero, and 0 otherwise.

Internal Financial Evidence: Peer Store Sales

To develop expectations at the daily store level, this study uses the contemporaneous daily sales of a peer store, *PEER*, as a control variable and relevant internal financial evidence. Allen et al. (1999) use all 29 available locations as peer stores. However, there are several hundred stores in this study. To choose the most relevant peer store,[5] this study assumes that a store with a sales pattern similar to that of the target store in a prior year tends to have a similar sales pattern in the current year. Using the daily store sales of the training period (i.e., fiscal year 2011), we correlate a target store's sales and candidate stores' sales and select the peer store with the most similar sales pattern, generating the highest correlation coefficient. The top 25 percent, the median, and the bottom 25 percent of the correlation coefficients between the target store sales and the corresponding peer store sales are 0.87, 0.83, and 0.79, respectively, which shows that target store sales are highly correlated with peer store.

Evaluation of the Performance of SAPs

Seeding Misstatements

Since we assume that the given financial data contains no misstatements, simulated misstatements are seeded into the testing dataset to evaluate the misstatement detection performance. Existing literature studying analytical procedures commonly uses this simulation approach, offering the benefits of controlled experiments (Vandervelde et al. 2008; Kogan et al. 2014; Rozario et al. 2023). In particular, 10 percent of these daily sales are seeded as misstatements because overstated revenue is widely used as a fraud scheme (Hogan, Rezaee, Riley, and Velury 2008).

Some prior studies, using relationships between one account balance and another to develop analytical models, assume either that both the dependent and independent accounts are misstated (i.e., coordinated misstatements) (Leitch and Chen 2003; Hoitash et al. 2006), or that only the dependent account is misstated (i.e., uncoordinated misstatements) (Chen and Leitch 1999). Although coordinated misstatements are considered common errors (Leitch and Chen 2003), they are harder to identify than uncoordinated misstatements because the predictor account's misstatement falls in the same direction as the dependent account's misstatement (Li 2016).

To test our second research question, we adopt both assumptions to modify internal financial data and examine whether the weather information plays a different role in enhancing misstatement detections depending on the misstatement seeding assumption. In the coordinated misstatement setting, we use both the actual store and peer store sales with misstatements and assume weak internal control. In the uncoordinated misstatement setting, we use the actual store sales with misstatements and actual peer store sales without misstatements and assume moderate or strong internal control. We expect that the weather information plays a more critical role in identifying misstatements precisely in the coordinated misstatement setting than in the uncoordinated setting because peer store sales with misstatements might contribute less toward detecting misstatements compared to the uncoordinated case. Accordingly, simulations are run for 365 prediction periods (365 days) for stores, for each misstatement assumption and prediction model (the benchmark and several weather models).

Out-of-Sample Forecasting

This study adopts an out-of-sample forecasting approach to test the performance of analytical procedures, following the previous studies' approaches (Hoitash et al. 2006; Vandervelde et al. 2008). The training set used to develop statistical models and the testing set used to assess misstatement detection performance are separated. This research employs modeling with a moving window to reflect possible structural changes in sales trends over certain periods (Kogan et al. 2014). The window size is 365 days. For instance, in order to estimate the 366th daily sales of a store, the sales of the

[5] We have considered other strategies to choose a peer store, including using the geographically closest one. However, it is important to note that geographically closest stores do not always share similar sales patterns with a target store, resulting in overall correlations between peer stores selected by geographic distance and target store sales that are lower than those achieved with the proposed approach. For instance, stores located in the countryside may have the closest one quite far away. Nevertheless, our conclusion may not be universally applicable to other cases as each firm has stores located differently. Therefore, the proper selection of peer stores may vary from case to case.

store and its peer store from the first to the 365th day are utilized as the training set. Then, the 367th daily sales of the store are estimated using the sales of the store from the second to the 366th day. Accordingly, a model is estimated for each observation of a store in the testing set based on its unique peer store.

Estimated Cost of Misclassification

After calculating forecasted daily sales, we identify misstatements if $(y_t - \hat{y}_t) > Z_{1\text{-}\alpha} \times (S_y/4)$, where y is daily sales, $Z_{1\text{-}\alpha}$ is the critical Z-value at the risk level, α, and S_y is the estimation period standard deviation of the series, y_t (Hoitash et al. 2006). Then, we identify false positive (FP) and false negative (FN) errors following prior analytical procedure research (e.g., Rozario et al. 2023). FP occurs when an analytical model points out a misstatement in a true account balance, and FN arises when the model fails to identify a materially misstated account balance.

The relationship between FP and FN is often a tradeoff. Normally, higher α levels lead to tighter prediction intervals and consequently result in fewer FNs and a larger number of FPs. FNs are often considered more expensive than FPs by auditors, but misstatements are less likely to occur. The measurement of misstatement detection performance is linked to this balance, determined by (1) the weight of FP and FN (i.e., a cost ratio between FPs and FNs), (2) the misstatement probability, i.e., the probability that a given observation contains an error or misstatement, and (3) the size of the prediction interval, determined by α, the significance level. Hence, as a measurement of misstatement detection performance, we calculate the estimated cost of misclassification (Beneish 1999), which considers FP, FN, the cost ratio, and the misstatement probability, P, and presents the estimated cost that auditors incur if the model does not correctly identify misstatements. Accordingly:

$$\begin{aligned} Estimated\ Cost\ of\ Misclassification = {} & Cost\ of\ FP \times P(FP \mid No\ Misstatement) \times P(No\ Misstatement) \\ & + Cost\ of\ FN \times P(FN \mid Misstatement) \times P(Misstatement) \end{aligned}$$

Since there are no guidelines for choosing an appropriate cost ratio or the misstatement probability, we assume, based on prior studies, that the probability of misstatement is 0.6 percent (Perols, Bowen, Zimmermann, and Samba 2017) following the findings of Bell and Carcello (2000) and the cost ratio between FP and FN is 1:30 (Beneish 1999). Accordingly, the estimated cost of misclassification for each store and model is formulated by (1 $\times$ P(FP | *No Misstatement*) $\times$ 99.4 percent) + (30 $\times$ P(FP | *Misstatement*) $\times$ 0.6 percent).

Also, we conduct a cost analysis to find the value for α that minimizes the cost of misclassification using the benchmark model, given the assumptions about the cost ratio and the misstatement probability. Specifically, the value for α is decided by conducting a cost analysis with the average FP and FN derived from the benchmark model over various α values. Table 1 shows the cost of misclassification estimated by the average FP and FN of the benchmark model over two misstatement seeding assumptions, multiple cost ratios, and α. We set the probability of misstatement at our primary setting of 0.6 percent and use four different cost ratios (1:10, 1:20, 1:30, and 1:40) to find the α that minimizes the cost of misclassification. In addition, the coordinated and uncoordinated misstatement seeding assumptions are considered separately to understand whether the results are consistent. Both Panels A and B show that when $\alpha = 0.05$, misclassification costs tend to be minimized regardless of various assumptions. Accordingly, $\alpha = 0.05$ is used for assessing the performance of misclassification detection.

To statistically test whether the estimated cost of misclassification derived from the weather model is smaller (larger) than the estimated cost of misclassification derived from the benchmark model, we adopt the nonparametric right-sided (left-sided) Wilcoxon Signed Rank Test, used to compare two models.[6]

IV. RESULTS

Descriptive Statistics

Table 2 reports the summary statistics for weather indicators. The mean values of *AT*, *HDD*, and *CDD* are approximately 59.60°F, 9.81°F, and 5.01°F, respectively. The mean value of *PRECIP* is 0.28. Table 3 illustrates the average correlation between store-level daily sales and weather indicators, calculated using correlation coefficients between individual store sales and corresponding weather indicators. The first column of Table 3 shows the overall correlation

[6] Because the object of this study is to test whether weather information offers incremental values in misstatement detection, we use the Wilcoxon Signed Rank test to compare two models and equally weight individual results. Whether the weather models outperform the benchmark models under conditions, like when the benchmark models barely identify misstatements, is not our research object, although this setting can offer strong evidence suggesting the use of weather information in auditing. We do not argue that weather information is particularly valuable when the benchmark models barely identify misstatements because we do not expect weather conditions to be the only influence the benchmark models do not contain.

TABLE 1

Cost of Misclassification Analysis for Store-Level Daily Sales Revenue

Panel A: Coordinated Misstatements

			Cost of FP to Cost of FN Ratios			
α	Average FP	Average FN	1:10	1:20	1:30	1:40
0.50	0.512	0.486	0.538	0.567	0.596	0.626
0.45	0.490	0.507	0.517	0.548	0.578	0.609
0.40	0.467	0.527	0.496	0.527	0.559	0.591
0.35	0.444	0.548	0.474	0.507	0.540	0.573
0.30	0.420	0.570	0.452	0.486	0.520	0.554
0.25	0.395	0.594	0.428	0.464	0.500	0.535
0.20	0.367	0.619	0.402	0.439	0.476	0.513
0.15	0.336	0.648	0.373	0.412	0.451	0.490
0.10	0.299	0.682	0.338	0.379	0.420	0.461
0.05	0.250	0.729	**0.292**	**0.336**	**0.380**	**0.423**

Panel B: Uncoordinated Misstatements

			Cost of FP to Cost of FN Ratios			
α	Average FP	Average FN	1:10	1:20	1:30	1:40
0.50	0.512	0.276	0.525	0.542	0.559	0.575
0.45	0.490	0.294	0.505	0.522	0.540	0.558
0.40	0.467	0.314	0.483	0.502	0.521	0.540
0.35	0.444	0.334	0.461	0.481	0.501	0.521
0.30	0.420	0.356	0.439	0.460	0.482	0.503
0.25	0.395	0.379	0.415	0.438	0.461	0.484
0.20	0.367	0.406	0.389	0.414	0.438	0.462
0.15	0.336	0.438	0.360	0.387	0.413	0.439
0.10	0.299	0.478	0.326	0.355	0.383	0.412
0.05	0.250	0.537	**0.281**	**0.313**	**0.345**	**0.377**

Table 1 presents the cost of misclassification, which is modified by the probability of misstatement, the cost ratios between FP and FN, and α. The average FPs and FNs are derived from the benchmark model SARIMAX $(1, 0, 0)(1, 0, 0)_7$ with *PEER*. The average FN is derived based on the uncoordinated misstatement distribution assumption only because FPs are the same regardless of misstatement seeding assumptions. Cost values are comparable in the same column but are not comparable across columns. The lowest costs are shown in bold.

coefficients between store-level daily sales and weather indicators. On average, the correlation coefficients for *AT*, *HDD*, and *CDD* are 0.26, −0.30, and 0.12, respectively, showing that sales are likely to increase as the temperature increases. The correlation coefficient for *PRECIP* is −0.09. As expected, unfavorable weather conditions are negatively associated with daily store sales. Table 3, columns (2)–(5) reports the correlation coefficients of weather indicators by region: Northeast, Midwest, South, and West based on the U.S. Census Bureau designations.[7] The daily sales for stores located in the Northeast and Midwest areas are more highly associated with temperature-related weather indices *AT*, *HDD*, and *CDD* than in the South and West areas. By contrast, the daily sales for stores located in the South and West areas are more highly associated with unfavorable weather conditions *PRECIP*, than stores in the Northeast and the

[7] Northeast (Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont, Delaware, New Jersey, New York, and Pennsylvania), Midwest (Illinois, Indiana, Michigan, Ohio, Wisconsin, Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, and South Dakota), South (Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia, District of Columbia, West Virginia, Alabama, Kentucky, Mississippi, Tennessee, Arkansas, Louisiana, Oklahoma, and Texas), and West (Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, Wyoming, Alaska, California, Hawaii, Oregon, and Washington). This simply describes the information from U.S. Census Bureau and does not indicate store locations we use in this study.

TABLE 2

Summary Statistics of Weather Variables

Weather Indicator	Mean (1)	Std. Dev. (2)	25% (3)	75% (4)
AT	59.599	20.419	45.000	74.000
HDD	9.812	12.803	0.000	18.000
CDD	5.007	7.235	0.000	10.000
PRECIP	0.275	0.446	0.000	1.000

Table 2 reports the summary statistics of weather indicators corresponding to the daily store sales data in the fiscal years 2011 and 2012. Variables are defined in Appendix A.

TABLE 3

The Average Correlation between Daily Store Sales and Weather Indicators by Regions

		Region			
Weather Indicator	Overall (1)	Northeast (2)	Midwest (3)	South (4)	West (5)
AT	0.260	0.340	0.374	0.172	0.269
HDD	−0.300	−0.358	−0.395	−0.226	−0.317
CDD	0.121	0.149	0.181	0.089	0.102
PRECIP	−0.088	−0.090	−0.014	−0.113	−0.113

Table 3 reports the average correlation between weather indicators and daily store sales data in the fiscal years 2011 and 2012. Variables are defined in Appendix A.

Midwest areas. Based on this analysis, we conclude that weather indices are generally associated with daily store sales. However, the direction of associations varies depending on the nature of the weather indicators.

Tests of Research Questions

Table 4 shows the misstatement detection performance of the benchmark model, SARIMAX $(1, 0, 0)(1, 0, 0)_7$ with *PEER*, and the weather model, SARIMAX $(1, 0, 0)(1, 0, 0)_7$ with *PEER* and *AT*. These outcomes are calculated based on the assumption that the probability of misstatement is 0.6 percent, the cost ratio between FP and FN is 1:30, and $\alpha = 0.05$. Under the coordinated misstatement assumption where daily store sales and peer store sales are jointly misstated, on average, the FP is 25 percent for the benchmark model and 24 percent for the weather model. Similarly, the FN is 72.9 percent for the benchmark model and 71.8 percent for the weather model. These results indicate that the weather model decreases FP and FN by 1 percent. The cost of misstatements for the benchmark and weather models are 0.38 and 0.37, respectively. The Wilcoxon Signed Rank test shows that the cost of misstatement derived from the weather model is statistically lower than the benchmark model. On average, the cost of misclassification decreases by 3.16 percent by adding the weather indicator.

Using the uncoordinated misstatement assumption where only target store sales are misstated, the FN of the benchmark and the weather models are 53.7 percent and 54.5 percent, respectively,[8] which indicates that adding the weather indicator increases FN by 0.8 percent on average. As expected, the FNs under the coordinated misstatement assumption are higher than under the uncoordinated misstatement assumption, indicating that it is more difficult to detect misstatements when both the store and peer sales are misstated. The cost of misstatements for the benchmark and weather models are about 0.35 and 0.34, respectively. This suggests that with the addition of the weather indicator, the cost of misclassification decreases by 2.32 percent, which is 0.84 percent lower than the reduction observed when adding the weather indicator under the coordinated misstatement assumption. Consistent with the results of the coordinated

[8] Because misstatement seeding assumptions only affect FN, FPs are the same regardless of misstatement seeding assumptions.

TABLE 4

Cost of Misclassification Analysis for Store-Level Daily Sales Revenue

Misstatement Distribution Assumption	Benchmark Model			Weather Model			Superior Model
	FP	FN	Cost	FP	FN	Cost	
Coordinated	0.250	0.729	0.380	0.240	0.718	0.368	Weather***
Uncoordinated	0.250	0.537	0.345	0.240	0.545	0.337	Weather***

*, **, *** Indicate significance at the 0.1, 0.05, and 0.01 levels, respectively.

This table presents the misstatement detection outcomes such as cost of misclassification, FPs, and FNs rates derived from the benchmark model (SARIMAX $(1, 0, 0)(1, 0, 0)_7$ with *PEER*) and from the weather model (SARIMAX $(1, 0, 0)(1, 0, 0)_7$ with *PEER* and *AT*). The risk level $\alpha = 0.05$, probability of misstatement = 0.6 percent, and the cost ratio 1:30 are employed. After material misstatements (10 percent of sales) are seeded in store daily sales, FNs are identified based on two misstatement seeding assumptions—coordinated (both *SALE* and *PEER* are misstated) and uncoordinated misstatements (only *SALE* is misstated). A superior model is chosen based on the costs of two models by the right (left)-sided Wilcoxon Signed Rank Test, which tests the hypothesis that costs from the weather (benchmark) model are lower than costs from the benchmark (weather) model.

Variables are defined in Appendix A.

misstatement assumption, the Wilcoxon Signed Rank Test shows that the cost of misclassification derived from the weather model is lower than the benchmark model and the results are significant at the 0.01 level. Overall, these results show that the weather model outperforms the benchmark model regardless of misstatement seeding assumptions. However, when both target store and peer store sales are misstated, the usefulness of the weather indicator is more evident than when only target store sales are misstated.

As robustness tests, we examine whether the results are consistent if the cost ratio, α, or the probability of misstatements are modified. Table 5, Panel A shows the cost of misclassification analysis with the 1:50 cost ratio. Because only the cost ratio is modified, FN and FP are unchanged, but the costs of misclassifications increase. Similarly, as Table 5, Panel B shows, as the probability of misstatements increases from 0.6 percent to 3 percent, FN and FP are consistent, but the cost of misclassification increases. Finally, Table 5, Panel C reports the results derived based on the assumption that α is 0.5 and the probability of misstatement is 3 percent. As α increases from 0.05 to 0.5, the FPs increase, but the FNs decrease. Overall, even if different assumptions are applied, the results are consistent with the results from Table 4, indicating that the weather model outperforms the benchmark model.

Alternative weather indicators are tested. First, both *HDD* and *CDD* are used for the weather model. Second, *AT* and *PRECIP* are used for the weather model. As in Table 6, we assume that α is 0.05 and the probability of misstatements is 0.6 percent. As before, we compare the weather model and the benchmark model based on coordinated and uncoordinated misstatement assumptions.

As Table 6, Panel A shows, the FP of the weather model with (1) *HDD* and *CDD* and (2) *PRECIP* and *AT* are 24.6 percent and 24 percent, respectively, which is marginally lower than the weather model with *AT*. Under the coordinated misstatement assumption, the FN of the weather model with (1) *HDD* and *CDD* and (2) *PRECIP* and *AT* are 72.8 and 71.8 percent, respectively, which is slightly higher than or equal to the weather model with *AT*. This tendency is consistent with the uncoordinated misstatement assumption as Table 6, Panel B illustrates. These results show that the weather model with *AT* identifies misstatements more precisely than the weather model with *HDD* and *CDD*. Adding *PRECIP* to the weather model might not drastically increase the misstatement detection performance, but the cost of misclassification derived from the weather model with alternative weather indices is statistically significantly lower than the benchmark model, indicating that adding weather indices enhances the misstatement detection performance.

Additional Analyses

For additional analysis, we first examine the misstatement detection performance by region. This analysis provides the opportunity to examine whether the usefulness of weather information varies by region. As Table 7 shows, the cost of misclassification, FPs, and FNs are organized into four. The FP of the benchmark model is 0.215–0.261, but the FP of the weather model is 0.208–0.251. Under the coordinated (uncoordinated) misstatement assumption, the FN of the benchmark model is 0.719–0.764 (0.510–0.567), but the FN of the weather model is 0.703–0.756 (0.311–0.344). Stores

TABLE 5

Alternative Assumptions for Cost of Misclassification Analysis

Panel A: Cost Ratio = 1:50

Misstatement Distribution Assumption	Benchmark Model FP	Benchmark Model FN	Benchmark Model Cost	Weather Model FP	Weather Model FN	Weather Model Cost	Superior Model
Coordinated	0.250	0.729	0.467	0.240	0.718	0.454	Weather***
Uncoordinated	0.250	0.537	0.410	0.240	0.545	0.402	Weather***

Panel B: α = 0.05 and Probability of Misstatement = 3%

Misstatement Distribution Assumption	Benchmark Model FP	Benchmark Model FN	Benchmark Model Cost	Weather Model FP	Weather Model FN	Weather Model Cost	Superior Model
Coordinated	0.250	0.729	0.899	0.240	0.718	0.879	Weather***
Uncoordinated	0.250	0.537	0.726	0.240	0.545	0.723	Weather***

Panel C: α = 0.5 and Probability of Misstatement = 3%

Misstatement Distribution Assumption	Benchmark Model FP	Benchmark Model FN	Benchmark Model Cost	Weather Model FP	Weather Model FN	Weather Model Cost	Superior Model
Coordinated	0.512	0.486	0.934	0.502	0.467	0.907	Weather***
Uncoordinated	0.512	0.276	0.745	0.502	0.277	0.736	Weather***

*, **, *** Indicate significance at the 0.1, 0.05, and 0.01 levels, respectively.
This table presents the misstatement detection outcomes such as cost of misclassification, FPs, and FNs rates derived from the benchmark model (SARIMAX $(1, 0, 0)(1, 0, 0)_7$ with *PEER*) and from the weather model (SARIMAX $(1, 0, 0)(1, 0, 0)_7$ with *PEER* and *AT*). The risk level, probability of misstatement, and the cost ratio are modified. After material errors (10 percent of sales) are seeded in store daily sales, FN is identified based on two misstatement seeding assumptions—coordinated (both *SALE* and *PEER* are misstated) and uncoordinated misstatements only (*SALE* is misstated). A superior model is chosen based on the costs of two models by the right (left)-sided Wilcoxon Signed Rank Test, which tests the hypothesis that costs from the weather (benchmark) model are lower than costs from the benchmark (weather) model.
Variables are defined in Appendix A.

TABLE 6

Alternative Weather Indicator

Panel A: Coordinated Misstatements

Benchmark Model FP	Benchmark Model FN	Benchmark Model Cost	Weather Indicator	Weather Model FP	Weather Model FN	Weather Model Cost	Superior Model
0.250	0.729	0.380	HDD_t, CDD_t	0.246	0.728	0.376	Weather***
0.250	0.729	0.380	$PRECIP_t$, AT_t	0.240	0.718	0.368	Weather***

(continued on next page)

located in any region tend to benefit from using *AT*. Among the four regions, stores located in the Northeast benefit the most from adopting the weather indicator.

Customers purchasing behaviors might also be associated with weather conditions yesterday, tomorrow, or rapid changes in recent weather. Consequently, we test four additional weather variables: weather conditions yesterday (AT_{t-1}); weather conditions tomorrow (AT_{t+1}); changes from yesterday ($AT_t - AT_{t-1}$); and changes from tomorrow

TABLE 6 (continued)

Panel B: Uncoordinated Misstatements

Benchmark Model			Weather Model				
FP	FN	Cost	Weather Indicator	FP	FN	Cost	Superior Model
0.250	0.537	0.345	HDD_t, CDD_t	0.246	0.542	0.342	Weather***
0.250	0.537	0.345	$PRECIP_t$, AT_t	0.240	0.545	0.337	Weather***

*, **, *** Indicate significance at the 0.1, 0.05, and 0.01 levels, respectively.
This table presents the misstatement detection outcomes such as cost of misclassification, FPs, and FNs rates derived from the benchmark model (SARIMAX $(1, 0, 0)(1, 0, 0)_7$ with *PEER*) and the weather model (SARIMAX $(1, 0, 0)(1, 0, 0)_7$ with *PEER* and weather indicator(s)). The risk level $\alpha = 0.05$, probability of error = 0.6 percent, and the cost ratio 1:30 are employed. After material misstatements (10 percent of sales) are seeded in store daily sales, FNs are identified based on two misstatement seeding assumptions—coordinated (both *SALE* and *PEER* are misstated) and uncoordinated misstatements (only *SALE* is misstated). A superior model is chosen based on the costs of two models by the right (left)-sided Wilcoxon Signed Rank Test, which tests the hypothesis that costs from the weather (benchmark) model are lower than costs from the benchmark (weather) model.
Variables are defined in Appendix A.

TABLE 7

Cost of Misclassification Analysis for Store-Level Daily Sales by Regions

Panel A: Coordinated Misstatements

Region	Benchmark Model			Weather Model			
	FP	FN	Cost	FP	FN	Cost	Superior Model
Northeast	0.215	0.764	0.351	0.208	0.756	0.343	Weather***
Midwest	0.245	0.734	0.376	0.237	0.728	0.367	Weather***
South	0.260	0.719	0.388	0.246	0.707	0.372	Weather***
West	0.261	0.720	0.389	0.251	0.703	0.376	Weather***

Panel B: Uncoordinated Misstatements

Region	Benchmark Model			Weather Model			
	FP	FN	Cost	FP	FN	Cost	Superior Model
Northeast	0.215	0.567	0.316	0.208	0.572	0.310	Weather***
Midwest	0.245	0.564	0.345	0.237	0.571	0.338	Weather***
South	0.260	0.526	0.353	0.246	0.536	0.341	Weather***
West	0.261	0.510	0.351	0.251	0.516	0.342	Weather***

*, **, *** Indicate significance at the 0.1, 0.05, and 0.01 levels, respectively.
This table presents the misstatement detection outcomes such as cost of misclassification, FPs, and FNs rates derived from the benchmark model (SARIMAX $(1, 0, 0)(1, 0, 0)_7$ with *PEER*) and from the weather model (SARIMAX $(1, 0, 0)(1, 0, 0)_7$ with *PEER* and *AT*) grouped by region. The risk level $\alpha = 0.05$, probability of misstatement = 0.6 percent, and the cost ratio 1:30 are employed. After material misstatements (10 percent of sales) are seeded in store daily sales, FNs are identified based on two misstatement seeding assumptions—coordinated (both *SALE* and *PEER* are misstated) and uncoordinated misstatements (only *SALE* is misstated). A superior model is chosen based on the costs of two models by the right (left)-sided Wilcoxon Signed Rank Test, which tests the hypothesis that costs from the weather (benchmark) model are lower than costs from the benchmark (weather) model. Regions are defined based on U.S. Census Bureau as follows: Northeast (Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont, Delaware, New Jersey, New York, and Pennsylvania), Midwest (Illinois, Indiana, Michigan, Ohio, Wisconsin, Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, and South Dakota), South (Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia, District of Columbia, West Virginia, Alabama, Kentucky, Mississippi, Tennessee, Arkansas, Louisiana, Oklahoma, and Texas), and West (Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, Wyoming, Alaska, California, Hawaii, Oregon, and Washington). This simply describes the information from U.S. Census Bureau and does not indicate store locations we use in this study.
Variables are defined in Appendix A.

TABLE 8

Weather Yesterday and Tomorrow and Changes in Weather Matter

Panel A: Coordinated Misstatements

Benchmark Model			Weather Model				
FP	**FN**	**Cost**	**Weather Indicator**	**FP**	**FN**	**Cost**	**Superior Model**
0.250	0.729	0.380	AT_{t-1}	0.240	0.719	0.368	Weather***
0.250	0.729	0.380	AT_{t+1}	0.240	0.718	0.368	Weather***
0.250	0.729	0.380	AT_t, $(AT_{t-1} - AT_t)$	0.239	0.717	0.367	Weather***
0.250	0.729	0.380	AT_t, $(AT_{t+1} - AT_t)$	0.239	0.717	0.367	Weather***

Panel B: Uncoordinated Misstatements

Benchmark Model			Weather Model				
FP	**FN**	**Cost**	**Weather Indicator**	**FP**	**FN**	**Cost**	**Superior Model**
0.250	0.537	0.345	AT_{t-1}	0.240	0.546	0.337	Weather***
0.250	0.537	0.345	AT_{t+1}	0.239	0.546	0.336	Weather***
0.250	0.537	0.345	AT_t, $(AT_{t-1} - AT_t)$	0.240	0.545	0.337	Weather***
0.250	0.537	0.345	AT_t, $(AT_{t+1} - AT_t)$	0.240	0.545	0.337	Weather***

*, **, *** Indicate significance at the 0.1, 0.05, and 0.01 levels, respectively.
This table presents the misstatement detection outcomes such as cost of misclassification, FPs, and FNs rates derived from the benchmark model (SARIMAX $(1, 0, 0)(1, 0, 0)_7$ with *PEER*) and the weather model (SARIMAX $(1, 0, 0)(1, 0, 0)_7$ with *PEER* and weather indicator(s)). The risk level $\alpha = 0.05$, probability of misstatement = 0.6 percent, and the cost ratio 1:30 are employed. After material misstatements (10 percent of sales) are seeded in store daily sales, FNs are identified based on two misstatement seeding assumptions—coordinated (both *SALE* and *PEER* are misstated) and uncoordinated misstatements (only *SALE* is misstated). A superior model is chosen based on the costs of two models by the right (left)-sided Wilcoxon Signed Rank Test, which tests the hypothesis that costs from the weather (benchmark) model are lower than costs from the benchmark (weather) model.
Variables are defined in Appendix A.

$(AT_{t+1} - AT_t)$. These weather change variables are tested with AT_t. As Table 8 reports, the results are consistent with the results shown in Table 4. These results might occur because weather conditions yesterday or tomorrow may not be significantly different from contemporaneous weather conditions.

V. CONCLUSION

The impact of advanced analytic technology has been highlighted by academia and audit practice. Under SAS 142, auditors can now access a wider range of external nonfinancial information than before. Yet, few studies have examined which measure can improve financial statement audits under which circumstances. By exploiting the relationship between weather indicators and sales, this study demonstrates that ENFMs derived from weather records enhance the performance of SAPs and are especially useful when internal financial information is not reliable, by using seeded error simulations into real-world store-level sales data.

For specific industries, weather records provide revenue-relevant information to assess financial performance. By considering the relationship between sales and weather, auditors can enhance analytical models to better detect material misstatements. Since weather information is timely and location-specific, it can also provide uniquely informative audit evidence. In addition, since weather information is external to the client and maintained by the regulatory agency, the reliability of audit evidence is enhanced.

This research attempts to fill a gap in the existing literature by determining how to develop analytical procedures that use external nonfinancial information and offer audit practitioners a chance to consider potential hurdles and opportunities inherent to adopting external nonfinancial data. Although prior studies have shown links between weather conditions and a firm's performance, it is uncertain when and how auditors should use such links for SAPs. This work contributes to audit practice and academia by illustrating whether, how, and when auditors can use weather information in SAPs for sales. Nevertheless, the findings of this study show that in order to utilize weather records, auditors may

American Accounting Association

need to expend extra effort in advance to examine the relationship between account balances and weather records by collecting, measuring, and incorporating this information. The varying and often limited degree to which practicing auditors rely on advanced statistical models may further constrain the utilization of the suggested models in audit practice.

The focus of this study is not to demonstrate a particular level of improvement in the results of SAPs, but rather to show the incremental value of utilizing weather records. We do not argue that weather indicators always provide incremental value in improving SAPs, that the suggested model is optimal, or that these weather indicators are relevant for auditors to consider in all contexts.

The results of this study suggest some further research opportunities. Because the association between weather and sales is complex, machine learning technology might be valuable for studying how the model could be changed by region, industry, or other additional factors. This study examines only one example of external data type and one account. It might be valuable to study the significance of new types of audit evidence by examining different accounts and data types. This study tests the value of weather records in SAPs; however, weather records can be utilized in other areas, such as enterprise risk assessment. In this case, different ways to measure weather records could be employed. This study links weather records with revenue, but depending on the firm, weather records might be useful in analyzing other accounts too.

Due to limited access to proprietary data, this study is constrained in scope. Further research might be needed into whether such associations between weather information and sales are modified depending on the location of stores (urban versus rural) or the modes of interaction with customers (online versus in-person). The influence of weather on customers' visits to stores may vary depending on the store locations. Also, we examine the retailer's store sales, which do not include online sales. We anticipate that the association between weather records and online sales may be weaker than the association between weather records and retail sales. However, it remains uncertain whether auditors can derive benefits from utilizing weather information for online sales because some prior studies insist that weather conditions affect customers' emotions, thereby influencing even online sales (e.g., Li et al. 2017). Further research on this question might be needed.

Finally, as we state in this manuscript, weather records are an example of ENFMs, and there are several potential ENFMs auditors might consider such as online news and consumer demographic data. Again, we do not argue that weather records are applicable to any case and might be better than other ENFMs. Future research examining how to utilize those various ENFMs in auditing and when they are beneficial for auditors might be needed.

REFERENCES

Allen, R. D., M. S. Beasley, and B. C. Branson. 1999. Improving analytical procedures: A case of using disaggregate multilocation data. *Auditing: A Journal of Practice & Theory* 18 (2): 128–142. https://doi.org/10.2308/aud.1999.18.2.128

American Institute of Certified Public Accountants (AICPA). 2012. Analytical procedure. In *AU-C Section 520*. New York, NY: AICPA.

American Institute of Certified Public Accountants (AICPA). 2018. *Agenda Item 5 Audit Evidence*. Durham, NC: AICPA.

American Institute of Certified Public Accountants (AICPA). 2020. *SAS 142: Audit Evidence*. New York, NY: AICPA.

Appelbaum, D., A. Kogan, and M. A. Vasarhelyi. 2017. Big Data and analytics in the modern audit engagement: Research needs. *Auditing: A Journal of Practice & Theory* 36 (4): 1–27. https://doi.org/10.2308/ajpt-51684

Arunraj, N. S., and D. Ahrens. 2015. A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting. *International Journal of Production Economics* 170: 321–335. https://doi.org/10.1016/j.ijpe.2015.09.039

Arunraj, N. S., D. Ahrens, and M. Fernandes. 2016. Application of SARIMAX model to forecast daily sales in food retail industry. *International Journal of Operations Research and Information Systems (IJORIS)* 7 (2): 1–21. https://doi.org/10.4018/IJORIS.2016040101

Badorf, F., and K. Hoberg. 2020. The impact of daily weather on retail sales: An empirical study in brick-and-mortar stores. *Journal of Retailing and Consumer Services* 52: 101921. https://doi.org/10.1016/j.jretconser.2019.101921

Barr-Pulliam, D., H. L. Brown-Liburd, and K.-A. Sanderson. 2022. The effects of the internal control opinion and use of audit data analytics on perceptions of audit quality, assurance, and auditor negligence. *Auditing: A Journal of Practice & Theory* 41 (1): 25–48. https://doi.org/10.2308/AJPT-19-064

Bell, T. B., and J. V. Carcello. 2000. A decision aid for assessing the likelihood of fraudulent financial reporting. *Auditing: A Journal of Practice & Theory* 19 (1): 169–184. https://doi.org/10.2308/aud.2000.19.1.169

Beneish, M. D. 1997. Detecting GAAP violation: Implications for assessing earnings management among firms with extreme financial performance. *Journal of Accounting and Public Policy* 16 (3): 271–309. https://doi.org/10.1016/S0278-4254(97)00023-9

Beneish, M. D. 1999. The detection of earnings manipulation. *Financial Analysts Journal* 55 (5): 24–36. https://doi.org/10.2469/faj.v55.n5.2296

Brazel, J. F., K. L. Jones, and D. F. Prawitt. 2014. Auditors' reactions to inconsistencies between financial and nonfinancial measures: The interactive effects of fraud risk assessment and a decision prompt. *Behavioral Research in Accounting* 26 (1): 131–156. https://doi.org/10.2308/bria-50630

Brown-Liburd, H., and M. A. Vasarhelyi. 2015. Big Data and audit evidence. *Journal of Emerging Technologies in Accounting* 12 (1): 1–16. https://doi.org/10.2308/jeta-10468

Bureau of Economic Analysis (BEA). 2020. *Gross Domestic Product, Fourth Quarter and Year 2019 (Third Estimate); Corporate Profits, Fourth Quarter and Year 201*. Washington, DC: BEA.

Chen, Y., and R. A. Leitch. 1999. An analysis of the relative power characteristics of analytical procedures. *Auditing: A Journal of Practice & Theory* 18 (2): 35–69. https://doi.org/10.2308/aud.1999.18.2.35

Christ, M. H., S. A. Emett, S. L. Summers, and D. A. Wood. 2021. Prepare for takeoff: Improving asset measurement and audit quality with drone-enabled inventory audit procedures. *Review of Accounting Studies* 26 (4): 1323–1343. https://doi.org/10.1007/s11142-020-09574-5

Christensen, B. E., R. J. Elder, and S. M. Glover. 2015. Behind the numbers: Insights into large audit firm sampling policies. *Accounting Horizons* 29 (1): 61–81. https://doi.org/10.2308/acch-50921

Cohen, J. R., G. Krishnamoorthy, and A. M. Wright. 2000. Evidence on the effect of financial and nonfinancial trends on analytical review. *Auditing: A Journal of Practice & Theory* 19 (1): 27–48. https://doi.org/10.2308/aud.2000.19.1.27

Cools, M., E. Moons, and G. Wets. 2009. Investigating the variability in daily traffic counts through use of ARIMAX and SARIMAX models: Assessing the effect of holidays on two site locations. *Transportation Research Record: Journal of the Transportation Research Board* 2136 (1): 57–66. https://doi.org/10.3141/2136-07

Forbes, S. J., and M. Lederman. 2010. Does vertical integration affect firm performance? Evidence from the airline industry. *The RAND Journal of Economics* 41 (4): 765–790. https://doi.org/10.1111/j.1756-2171.2010.00120.x

Glover, S. M., D. F. Prawitt, and M. S. Drake. 2015. Between a rock and a hard place: A path forward for using substantive analytical procedures in auditing large P&L accounts: Commentary and analysis. *Auditing: A Journal of Practice & Theory* 34 (3): 161–179. https://doi.org/10.2308/ajpt-50978

Griffith, E. E., J. S. Hammersley, and K. Kadous. 2015. Audits of complex estimates as verification of management numbers: How institutional pressures shape practice. *Contemporary Accounting Research* 32 (3): 833–863. https://doi.org/10.1111/1911-3846.12104

Hogan, C. E., Z. Rezaee, R. A. Riley, and U. K. Velury. 2008. Financial statement fraud: Insights from the academic literature. *Auditing: A Journal of Practice & Theory* 27 (2): 231–252. https://doi.org/10.2308/aud.2008.27.2.231

Hoitash, R., A. Kogan, and M. A. Vasarhelyi. 2006. Peer-based approach for analytical procedures. *Auditing: A Journal of Practice & Theory* 25 (2): 53–84. https://doi.org/10.2308/aud.2006.25.2.53

Hubbard, K. G., N. B. Guttman, J. You, and Z. Chen. 2007. An improved QC process for temperature in the daily cooperative weather observations. *Journal of Atmospheric and Oceanic Technology* 24 (2): 206–213. https://doi.org/10.1175/JTECH1963.1

Kapoor, M. 2020. *Big Four Invest Billions in Tech, Reshaping Their Identities*. New York, NY: Bloomberg.

Kirk, M., D. Stice, and H. Stice. 2022. The real effects of weather on firm performance, analyst forecasts, and stock prices. (Working paper). https://ssrn.com/abstract=4007090

Knechel, W. R. 1988. The effectiveness of statistical analytical review as a substantive auditing procedure: A simulation analysis. *The Accounting Review* 63: 74–95.

Kogan, A., B. W. Mayhew, and M. A. Vasarhelyi. 2019. Audit data analytics research—An application of design science methodology. *Accounting Horizons* 33 (3): 69–73. https://doi.org/10.2308/acch-52459

Kogan, A., M. G. Alles, M. A. Vasarhelyi, and J. Wu. 2014. Design and evaluation of a continuous data level auditing system. *Auditing: A Journal of Practice & Theory* 33 (4): 221–245. https://doi.org/10.2308/ajpt-50844

Koskivaara, E. 2004. Artificial neural networks in analytical review procedures. *Managerial Auditing Journal* 19 (2): 191–223. https://doi.org/10.1108/02686900410517821

Leitch, R. A., and Y. Chen. 2003. The effectiveness of expectation models in recognizing error patterns and generating and eliminating hypotheses while conducting analytical procedures. *Auditing: A Journal of Practice & Theory* 22 (2): 147–170. https://doi.org/10.2308/aud.2003.22.2.147

Lev, B. 1980. On the use of index models in analytical reviews by auditors. *Journal of Accounting Research* 18 (2): 524–550. https://doi.org/10.2307/2490591

Li, C., X. Luo, C. Zhang, and X. Wang. 2017. Sunny, rainy, and cloudy with a chance of mobile promotion effectiveness. *Marketing Science* 36 (5): 762–779. https://doi.org/10.1287/mksc.2017.1044

Li, P. 2016. Analytics with exception prioritization, consumer search volume, and social capital. Doctoral dissertation, Rutgers, The State University of New Jersey, Newark.

Maletta, M., and A. Wright. 1996. Audit evidence planning: An examination of industry error characteristics. *Auditing: A Journal of Practice & Theory* 15 (1): 71.

National Association of State Boards of Accountancy (NASBA). 2021. AICPA and NASBA release CPA evolution model curriculum. https://nasba.org/blog/2021/06/15/cpa-evolution-model-curriculum/

Omura, T., and R. Willett. 2006. Using automated equilibrium correction modelling in analytic review. *Managerial Auditing Journal* 21 (2): 207–223. https://doi.org/10.1108/02686900610639329

Osczevski, R., and M. Bluestein. 2005. The new wind chill equivalent temperature chart. *Bulletin of the American Meteorological Society* 86 (10): 1453–1458. https://doi.org/10.1175/BAMS-86-10-1453

Peak Resorts, Inc. 2015. Form 10-K. https://www.sec.gov/Archives/edgar/data/1517401/000151740115000023/skis-20150430x10k.htm

Peffers, K., T. Tuunanen, M. A. Rothenberger, and S. Chatterjee. 2007. A design science research methodology for information systems research. *Journal of Management Information Systems* 24 (3): 45–77. https://doi.org/10.2753/MIS0742-1222240302

Pérez-González, F., and H. Yun. 2013. Risk management and firm value: Evidence from weather derivatives. *The Journal of Finance* 68 (5): 2143–2176. https://doi.org/10.1111/jofi.12061

Perols, J. L., R. M. Bowen, C. Zimmermann, and B. Samba. 2017. Finding needles in a haystack: Using data analytics to improve fraud prediction. *The Accounting Review* 92 (2): 221–245. https://doi.org/10.2308/accr-51562

Public Company Accounting Oversight Board (PCAOB). 2013. Report on the PCAOB's 2007-2010 Inspections of Domestic Firm That Audit 100 or Fewer Public Companies. Release No. 2013-001. Washington, DC: PCAOB.

Public Company Accounting Oversight Board (PCAOB). 2014. Staff Audit Practice Alert No 12. Washington, DC: PCAOB.

Rose, N., and L. Dolega. 2022. It's the weather: Quantifying the impact of weather on retail sales. *Applied Spatial Analysis and Policy* 15 (1): 189–214. https://doi.org/10.1007/s12061-021-09397-0

Rothfusz, L. P. 1990. The heat index equation (or, more than you ever wanted to know about heat index). National Weather Service. https://www.weather.gov/media/ffc/ta_htindx.PDF

Rozario, A. M., M. A. Vasarhelyi, and T. D. Wang. 2023. On the use of consumer tweets to assess the risk of misstated revenue in consumer-facing industries: Evidence from analytical procedures. *Auditing: A Journal of Practice & Theory* 42 (2): 207–229. https://doi.org/10.2308/AJPT-2020-078

Starr-McCluer, M. 2000. The effects of weather on retail sales. In *Federal Reserve Board Discuss Series: Divisions of Research & Statistics and Monetary Affairs*. Washington, DC: Federal Reserve Board.

Štulec, I., K. Petljak, and D. Naletina. 2019. Weather impact on retail sales: How can weather derivatives help with adverse weather deviations? *Journal of Retailing and Consumer Services* 49: 1–10. https://doi.org/10.1016/j.jretconser.2019.02.025

Trompeter, G., and A. Wright. 2010. The world has changed—Have analytical procedure practices? *Contemporary Accounting Research* 27 (2): 669–700. https://doi.org/10.1111/j.1911-3846.2010.01021.x

Trotman, K. T., and W. F. Wright. 2012. Triangulation of audit evidence in fraud risk assessments. *Accounting, Organizations and Society* 37 (1): 41–53. https://doi.org/10.1016/j.aos.2011.11.003

Tysiac, K. 2020. New audit evidence standard recognizes evolving nature of business. *Journal of Accountancy* (July 9). https://www.journalofaccountancy.com/news/2020/jul/audit-evidence-standard-enables-technology-aided-audit-procedures.html

Vandervelde, S. D., Y. Chen, and R. A. Leitch. 2008. Auditors' cross-sectional and temporal analysis of account relations in identifying financial statement misstatements. *Auditing: A Journal of Practice & Theory* 27 (2): 79–107. https://doi.org/10.2308/aud.2008.27.2.79

Wright, A., and R. H. Ashton. 1989. Identifying audit adjustments with attention-directing procedures. *The Accounting Review* 64 (4): 710–728.

Yoon, K., and T. Pearce. 2021. Can substantive analytical procedures with data and data analytics replace sampling as tests of details? *Journal of Emerging Technologies in Accounting* 18 (2): 185–199. https://doi.org/10.2308/JETA-19-03-23-10

Yoon, K., L. Hoogduin, and L. Zhang. 2015. Big Data as complementary audit evidence. *Accounting Horizons* 29 (2): 431–438. https://doi.org/10.2308/acch-51076

Yoon, K., Y. Liu, T. Chiu, and M. A. Vasarhelyi. 2021. Design and evaluation of an advanced continuous data level auditing system: A three-layer structure. *International Journal of Accounting Information Systems* 42: 100524. https://doi.org/10.1016/j.accinf.2021.100524

APPENDIX A

Variable Definitions

Variables	Definition	Source
Dependent Variable		
SALE	Daily store-level sales.	Proprietary data from a global accounting firm[a]
Weather Indicator		
AT	Apparent temperature, equal to Heat Index (*HI*) if the average temperature $> 80°$F; the average temperature if $50°\text{F} \leq$ the average temperature $\leq 80°$F; Wind Chill Index (*WCI*) if the average temperature $< 50°$F.	Wunderground
HDD	Heating degree days: the difference between 65°F and the average daily temperature if the average daily temperature is below 65°F.	
CDD	Cooling degree days: the difference between 65°F and the average daily temperature if the average daily temperature exceeds 65°F.	
PRECIP	A dummy variable is coded as 1 if the sum of the daily precipitation is greater than zero inches, and 0 otherwise.	
Control Variable		
PEER	Peer store daily sales.	Proprietary data

[a] We obtain access to the data under a nondisclosure agreement. Because it is proprietary, we are not able to disclose its descriptive statistics and any information that might indicate the target firm's identity, such as the number of stores we use in this study, detailed information regarding goods and services the firm offers, and others.

APPENDIX B

Heat Index[9]

Heat Index (*HI*) is developed as follows:

$$HI = -42.379 + 2.04901523 \times T + 10.14333127 \times RH - 0.22475541 \times T \times RH 0.00683783 \times T^2 - 0.05481717 \times RH^2 + 0.00122874 \times T^2 \times RH + 0.00085282 \times T \times RH^2 - 0.00000199 \times T^2 RH^2$$

where:

$T =$ air temperature, °F; and
$RH =$ relative humidity in percent.

If the *RH* is less than 13 percent and the *T* is between 80 and 112°F, then the following adjustment is subtracted from *HI*:

$$Adjustment = \frac{(13 - RH)}{4} \times \sqrt{\frac{17 - |T - 95|}{17}}$$

If the *RH* is greater than 85 percent and *T* is between 80 and 87°F, then he following adjustment is added to *HI* is adjusted:

$$Adjustment = \frac{(RH - 85)}{10} \times \frac{(87 - T)}{5}$$

If *T* is below 90°F, then *HI* is formulated as follows:

$$HI = 0.5 \times (T + 61) + ((T\text{–}68) \times 1.2) + (RH \times 0.094)$$

[9] http://www.wpc.ncep.noaa.gov/html/heatindex_equation.shtml

APPENDIX C

Alternative Statistical Models

Panel A: SARIMAX (*1, 0, 1*)(*1, 0, 0*)$_7$

Misstatement Distribution Assumption	Benchmark Model			Weather Model			Superior Model
	FP	FN	Cost	FP	FN	Cost	
Coordinated	0.228	0.750	0.362	0.218	0.736	0.350	Weather***
Uncoordinated	0.228	0.557	0.327	0.218	0.564	0.319	Weather***

Panel B: SARIMAX (*1, 1, 1*)(*0, 1, 1*)$_7$

Misstatement Distribution Assumption	Benchmark Model			Weather Model			Superior Model
	FP	FN	Cost	FP	FN	Cost	
Coordinated	0.216	0.760	0.353	0.208	0.742	0.342	Weather***
Uncoordinated	0.216	0.566	0.318	0.208	0.573	0.311	Weather***

*, **, *** Indicate significance at the 0.1, 0.05, and 0.01 levels, respectively.
This table presents the misstatement detection outcomes such as cost of misclassification, FPs, and FNs rates derived from the benchmark model with *PEER* and from the weather model with *PEER* and *AT*. The risk level $\alpha = 0.05$, probability of misstatement = 0.6 percent, and the cost ratio 1:30 are employed. After material misstatements (10 percent of sales) are seeded in store daily sales, FNs are identified based on two misstatement seeding assumptions—coordinated (both *SALE* and *PEER* are misstated) and uncoordinated misstatements (only *SALE* is misstated). A superior model is chosen based on the costs of two models by the right (left)-sided Wilcoxon Signed Rank Test, which tests the hypothesis that costs from the weather (benchmark) model are lower than costs from the benchmark (weather) model.
Variables are defined in Appendix A.