Advances in Accounting xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Advances in Accounting

journal homepage: www.elsevier.com/locate/adiac

Miscodings in Compustat's auditor variable: issues, identification, and correction

Steven Utke

University of Connecticut, 2100 Hillside Rd., Unit 1041A, Storrs, CT 06269, United States

ARTICLE INFO

Keywords:
Auditor changes
Auditor tenure
Industry specialization
Big 4
Data integrity
Compustat

JEL codes:
M40

ABSTRACT

Using an easily implementable methodology for identifying potential data errors, I identify and correct cases where Compustat miscodes its auditor variable. In this paper, I present the methodology and provide SAS code that implements the methodology, enabling researchers to easily identify and correct auditor variable miscodings. Further, I provide a list of corrections for a sample of Compustat firms from 2001 to 2014. Auditor variable miscodings have implications for both audit-specific research as well as general capital markets research. I find that some of the miscodings arise from the fact that, following an auditor change, the previous auditor's report remains in a firm's 10-K, and Compustat occasionally codes the previous auditor as the current auditor. Aside from identifying and correcting miscodings, I also find that a non-zero number of firms change to a new auditor and then, after only one year with the new auditor, switch back to the prior auditor.

1. Introduction

Compustat serves as a major tool in performing archival research in accounting and finance. While most research uses Compustat as is, recent research identifies limitations of the Compustat data (e.g., Boritz & No, 2013; Casey, Gao, Kirschenheiter, Li, & Pandit, 2016; Chychyla & Kogan, 2014; Heitzman & Lester, 2018; Keil, 2017; Mills, Newberry, & Novack, 2003). For example, Mills et al. (2003) note that Compustat sometimes miscodes net operating loss carryforwards (NOLs) as zero or missing when a disclosed value exists. Casey et al. (2016) establish an overall process for filling in missing Compustat values with an appropriate value, calculated from other information, or with zeros when appropriate. While Casey et al. (2016) note that Compustat goes through an extensive data validation process, this existing research shows that miscoding occasionally occurs in the Compustat data. The potential for miscoding is likely highest for information that only appears in footnotes rather than on the face of the financial statements (e.g., the NOL disclosures identified by Mills et al., 2003).

The auditor variable (Compustat variable AU) is one such variable subject to potential miscoding in Compustat. The auditor variable is important in a wide range of capital markets research, but especially in audit research focused on the Big N/non-Big N distinction (e.g., DeFond, Erkens, & Zhang, 2017; Lawrence, Minutti-Meza, & Zhang, 2011), industry specialization (e.g., Gaver & Utke, 2018; Minutti-Meza, 2013), auditor tenure (e.g., Gul, Fung, & Jaggi, 2009; Myers, Myers, & Omer, 2003), auditor changes (e.g., DeFond & Subramanyam, 1998), and related areas. While some recent audit studies rely on Audit Analytics instead of Compustat, Compustat remains heavily used in both general capital markets research and audit research. For example, of the audit studies cited above, only DeFond et al. (2017) use Audit Analytics as their main data source. Further, Audit Analytics only covers years after 2000, making Compustat necessary for studies of earlier time periods, which continue to be conducted (e.g., Choi, Kim, & Raman, 2017; Jiang, Wang, & Wang, 2018; Kraft, Vashishtha, & Venkatachalam, 2018). Compustat is also necessary for auditor tenure studies, which generally require historical data over many decades in order to compute auditor tenure (see, e.g., Singer & Zhang, 2018). More broadly, Compustat and its auditor variable are commonly used in general capital markets studies not focusing specifically on auditors (e.g., Huang, Jennings, & Yu, 2017). Finally, there are issues merging Audit Analytics and Compustat; as such, relying on Audit Analytics may be problematic in, for example, industry specialization studies that require Compustat variables such as sales, assets, or market value to calculate industry specialization.[1]

Because Compustat continues to play a large role in archival capital markets and audit research, miscoding in the auditor variable can affect

E-mail address: sutke@uconn.edu.

[1] For example, Reichelt and Wang (2010) begin with Audit Analytics data and lose over 20% of their sample when merging with Compustat. Compustat and Audit Analytics do not merge well for two reasons: a) Compustat assigns a header (i.e., current) CIK to each firm-year observation, whereas Audit Analytics uses the historical CIK and b) some Compustat CIKs have errors, with the most notable CIK error likely being for Schering-Plough (gvkey 009459). Full discussion of these issues is beyond the scope of this paper.

https://doi.org/10.1016/j.adiac.2018.09.002
Received 23 August 2018; Accepted 9 September 2018
0882-6110/

studies of Big N effects, industry specialization, auditor tenure, and auditor changes, among others. For example, Gaver and Utke (2018) examine the relation between the length of time an auditor serves as an industry specialist and audit quality. Without adjusting for miscodings, identification of the industry specialist and *changes* in the industry specialist may be driven by miscoding rather than actual auditor attributes. Failing to account for miscoding would add noise to their industry specialization measures (because specialists would be misclassified as non-specialists, and vice versa), lead to incorrect identification of changes in an industry's specialist auditor, and lead to incorrect calculations of the time a specialist auditor has served as a specialist in a given industry (i.e., the specialist's tenure).[2] Similar issues could arise in other studies of auditor related effects.

In this paper, I discuss the issues that can lead to miscoding in Compustat's auditor variable. I then present a simple methodology for identifying these miscoded auditors. I provide SAS code that allows researchers to implement the methodology quickly and easily in order to correct their data. I also provide SAS code to implement the corrections applicable to a sample in which I previously identified miscodings, allowing future researchers to easily make these corrections. I briefly discuss an example miscoding and show its potential implications for audit research. In implementing my methodology, I also find that a non-zero number of firms change to a new auditor and then, after only one year with the new auditor, switch back to the prior auditor. These firms may warrant special consideration in future audit research.

2. Miscoding issues and identification

According to Casey et al. (2016), Compustat's data collection efforts include extraction of information from the financial statement notes. While the specific methodology behind this extraction is unclear (e.g., the extent to which the extraction is manual versus automated), the auditor variable has a high potential for miscoding in some circumstances. Specifically, in the two years following an auditor change, a firm's former auditor must include its prior year audit reports in the firm's financial statements. The prior auditor's report is often presented immediately after the current auditor's report and appears nearly identical to the current report, except for the financial statement dates referenced in the report and the auditor's signature date. As such, either a manual or automated data collection system could have difficulty distinguishing between these two reports, leading to miscoding of the previous auditor as the current auditor. See Appendix C for an example of this presentation that resulted in a Compustat miscoding.

Beginning with this understanding of what potentially causes miscoding in the auditor variable, it is relatively easy to identify cases with possible miscoding. Specifically, a miscoding is possible (but not certain) in cases where, according to the Compustat data, an auditor change occurs from year $t-1$ to year t, and again from year t to year $t+1$, with the new (year $t+1$) auditor equal to the original (year $t-1$) auditor. Said differently, potential miscodings exist when the year $t-1$ and year $t+1$ auditors are the same, but the year t auditor differs. I refer to these situations as “switch backs,” and switch backs encompass miscodings as well as firms actually switching back to their former auditor. Note that an auditor change is not required in order to have a miscoding (e.g., miscoding may result from a simple data entry issue) and the methodology I develop also identifies these situations. Section 3 discusses the frequency of miscodings caused by auditor changes versus miscodings with no apparent cause. Appendix A presents SAS code to identify switch backs – including both miscodings and actual switch backs.

As mentioned above, not all switch backs represent miscodings. For example, in the early 2000s, several firms switch to Arthur Andersen as their new auditor, and then switch back to their prior auditor following Andersen's collapse. Also in the early 2000s, some firms switch to a new auditor, receive an internal control weakness opinion from the new auditor, and then switch back to the prior auditor. Preliminary analyses of these events suggested that they occur infrequently, limiting the ability to study these events separately. However, future studies that examine auditor changes may benefit from performing additional analyses on switch back firms.

3. Data and corrections

In order to identify potential miscoding in the Compustat auditor variable, I obtain all Compustat firm-years from 2000 to 2015. Because I require a one year lag and a one year lead, this effectively limits my corrections to the years 2001 to 2014.[3] I require that the firm-year report both total assets and sales of greater than zero. This yields 130,905 observations. I focus on firms likely to have U.S. auditors, so I retain only firm-years located in the US. I also retain only firm-years with a Big 4 auditor (plus Arthur Andersen) in the current or prior year, as the Big 4 audit the substantial majority (approximately 95%) of the market value of firms during my sample period. After imposing these requirements, the final sample contains 66,365 firm-years. Table 1 presents the sample selection process.[4]

From this sample, I apply the identification process discussed in Section 2 using the code in Appendix A. I identify 322 potentially miscoded auditors. This exceeds the actual number of miscodings identified below for two reasons. First, as discussed above, some of these observations involve real auditor switch backs. Second, for firms with miscodings associated with auditor changes, the identification process generally identifies both the miscoded year and the previous year (e.g., the year of the actual auditor change). To determine the proper characterization of the 322 potentially miscoded auditors, I obtain 10-Ks from the SEC's EDGAR database for the relevant firm years and manually review the auditor information.[5] Out of the 322 observations, I identify 98 miscodings where Compustat used the prior auditor as the current year auditor, and 125 cases that simply involved miscoding with no clear explanation. Note that before implementing my methodology in this sample, I separately identified seven miscodings (only two of which involved Compustat using the prior auditor as the current auditor). Thus, out of the 66,365 observations examined, 230 (0.35%) miscodings exist.

While the miscoding rate is low in the full sample, there are relatively fewer auditor changes each year, so miscodings make up a larger percentage of auditor change observations than of total observations. Further, miscodings occur on significant firms such as Verizon and therefore can be meaningful, especially when determining an auditor's industry specialization or tenure. Specifically for Verizon, an auditor variable miscoding occurred in 2012. Verizon had been with its auditor, Ernst and Young (EY), since 2000 but was miscoded as using KPMG in 2012. Table 2 presents the effect of this miscoding on various commonly calculated auditor variables. Without correcting for the miscoding in 2012, auditor tenure for 2012 through 2015 is misstated as: 1,

[2] See the discussion of Table 2 below for further details. Gaver and Utke's (2018) analyses account for all auditor variable miscodings discussed in this paper.

[3] While hand collecting the auditor data for the switch backs identified using the process described in this paper, I also identified several corrections for 2015. These corrections are included in Appendix B.

[4] As noted earlier, Audit Analytics and Compustat do not merge perfectly. To show this, I merge Audit Analytics into my Compustat sample, without adjusting for the header versus historical CIK issue or the CIK errors discussed earlier. Approximately 9% of the Compustat sample does not merge with Audit Analytics. Thus my methodology to correct Compustat miscodings remains useful despite the availability of auditor information in Audit Analytics.

[5] EDGAR generally dates back only to 1996. However, 10-Ks for firm years between 1988 and 1996 are often available electronically from Thomson ONE. Earlier 10-Ks are available directly from the SEC for a small fee.

Table 1
Sample selection.

Data restrictions	N
Compustat firms from 2000 to 2015 with non-missing with total assets and sales > 0	130,905
Less:	
Non-U.S. firms	(32,922)
Firm-years not audited by Big 4 in current or prior year	(31,618)
Sample size	66,365

I identify U.S. firms using the following procedure. Because Compustat generally treats location (LOC) as a header variable, I first merge in the historical location (HLOC). However, this variable is only available beginning in 2007. For all firm-years with HLOC available, I backfill prior years using the earliest available HLOC. If a firm-year still does not have a location (e.g., because the firm only existed prior to 2007), I backfill the location using LOC.

These miscodings have potential implications for audit research, as well as other capital markets research that relies on auditor information. I present an easily implementable methodology for identifying these miscodings so that researchers can correct them. I also provide a number of corrections for miscodings that occurred from 2001 to 2014 for a sample of firms with Big 4 auditors in either the current or prior year. I provide SAS code for implementing the methodology and corrections. Researchers should implement these corrections when using Compustat auditor data. I also find that some firms change auditors and then switch back to their previous auditor within a short period of time. Researchers specifically studying auditor changes may benefit from separately examine these cases.

Table 2
Example of effects of auditor miscoding on auditor variables - Verizon.

Year	2-Digit SIC	Using "as reported" values						Using corrected values					
		Verizon's Auditor	Auditor Tenure	Industry Specialist (30%)	Industry Specialist (Leader by 10%)	Specialist Tenure (30%)	Specialist Tenure (Leader by 10%)	Verizon's Auditor	Auditor Tenure	Industry Specialist (30%)	Industry Specialist (Leader by 10%)	Specialist Tenure (30%)	Specialist Tenure (Leader by 10%)
2000	48	EY	1	EY	EY	1	1	EY	1	EY	EY	1	1
2001	48	EY	2	EY	EY	2	2	EY	2	EY	EY	2	2
2002	48	EY	3	EY	EY	3	3	EY	3	EY	EY	3	3
2003	48	EY	4	EY	EY	4	4	EY	4	EY	EY	4	4
2004	48	EY	5	EY	None	5	0	EY	5	EY	None	5	0
2005	48	EY	6	EY	EY	6	1	EY	6	EY	EY	6	1
2006	48	EY	7	EY	EY	7	2	EY	7	EY	EY	7	2
2007	48	EY	8	EY	EY	8	3	EY	8	EY	EY	8	3
2008	48	EY	9	EY	EY	9	4	EY	9	EY	EY	9	4
2009	48	EY	10	EY	EY	10	5	EY	10	EY	EY	10	5
2010	48	EY	11	EY	EY	11	6	EY	11	EY	EY	11	6
2011	48	EY	12	EY	EY	12	7	EY	12	EY	EY	12	7
2012	48	KPMG	1	KPMG	None	1	0	EY	13	EY	EY	13	8
2013	48	EY	1	EY	EY	1	1	EY	14	EY	EY	14	9
2014	48	EY	2	EY	EY	2	2	EY	15	EY	EY	15	10
2015	48	EY	3	EY	EY	3	3	EY	16	EY	EY	16	11

1, 2, and 3 years, when the correct values are 13, 14, 15, and 16 years. Regarding U.S. national industry specialization, Verizon's actual auditor (EY) is the specialist for 2-digit SIC 48 for 2011, 2012, and 2013, measured by both the 30% of market share cut-off and the market leader by 10% cutoff (Reichelt & Wang, 2010), where market share is measured by client sales. When Verizon's auditor is miscoded to KPMG in 2012, KPMG appears to be the 2012 market share leader using the 30% cutoff, but there is no specialist identified using the leader by 10% measure. Without correcting for the miscoding, specialist tenure is miscalculated as well.[6] Importantly, the single year miscoding affects the auditor tenure calculations for *all* future years.

Appendix B provides SAS code for correcting the miscodings identified in the sample detailed in Table 1. For reference, I also provide information on the firm-years that appear to switch back to their previous auditor (32 observations). After identifying and correcting observations using the procedures discussed here, I suggest researchers iterate through the process as new potential miscodings may arise after correcting the original miscodings.

4. Conclusion

This paper identifies miscoding in Compustat's auditor variable.

[6] Because the miscoding was from EY to KPMG, the Big 4 designation would be correct across all years in this case. However, this is not always true as miscodings can involve auditors in different "tiers."

Data availability

Data used in this study are available from public sources identified in the paper.

Declarations of conflicts of interest

None.

Funding

Author acknowledges the support of the University of Connecticut School of Business.

Acknowledgment

This paper has benefited from helpful comments from Tom Adams, Jenny Gaver, Youree Kim, and Quinn Swanquist.

Appendix A. Supplementary data: SAS code and example auditor's report

Supplementary data to this article can be found online at https://doi.org/10.1016/j.adiac.2018.09.002.

References

Boritz, J. E., & No, W. G. (2013). *The quality of interactive data: XBRL versus Compustat, Yahoo Finance, and Google Finance.* (Working paper).

Casey, R. J., Gao, F., Kirschenheiter, M. T., Li, S., & Pandit, S. (2016). Do Compustat financial statement data articulate? *Journal of Financial Reporting, 1*(1), 37–59.

Choi, J.-H., Kim, S., & Raman, K. K. (2017). Did the 1998 merger of Price Waterhouse and Coopers & Lybrand increase audit quality? *Contemporary Accounting Research, 34*(2), 1071–1102.

Chychyla, R., & Kogan, A. (2014). *Does Compustat data standardization improve bankruptcy prediction models?* Working paperRutgers University.

DeFond, M., Erkens, D. H., & Zhang, J. (2017). Do client characteristics really drive the Big N audit quality effect? New evidence from propensity score matching. *Management Science, 63*(11), 3628–3649.

DeFond, M. L., & Subramanyam, K. R. (1998). Auditor changes and discretionary accruals. *Journal of Accounting and Economics, 25*, 35–67.

Gaver, J. J., & Utke, S. (2018). Audit quality and specialist tenure. *The Accounting Review* (forthcoming).

Gul, F. A., Fung, S. Y. K., & Jaggi, B. (2009). Earnings quality: Some evidence on the role of auditor tenure and auditors' industry expertise. *Journal of Accounting and Economics, 47*, 265–287.

Heitzman, S., & Lester, R. (2018). *Net operating loss carryforwards and corporate financial policies.* (Working paper).

Huang, Y., Jennings, R., & Yu, Y. (2017). Product market competition and managerial disclosure of earnings forecasts: Evidence from import tariff rate reductions. *The Accounting Review, 92*(3), 185–207.

Jiang, J. X., Wang, I. Y., & Wang, K. P. (2018). Big N auditors and audit quality: New evidence from quasi-experiments. *The Accounting Review* (forthcoming).

Keil, J. (2017). The trouble with approximating industry concentration from Compustat. *Journal of Corporate Finance, 45*, 467–479.

Kraft, A. G., Vashishtha, R., & Venkatachalam, M. (2018). Frequent financial reporting and managerial myopia. *The Accounting Review, 93*(2), 249–275.

Lawrence, A., Minutti-Meza, M., & Zhang, P. (2011). Can Big 4 versus non-Big 4 differences in audit-quality proxies be attributed to client characteristics? *The Accounting Review, 86*(1), 259–286.

Mills, L. F., Newberry, K. J., & Novack, G. F. (2003). How well do Compustat NOL data identify firms with U.S. tax return loss carryovers? *Journal of American Taxation Association, 25*(2), 1–17.

Minutti-Meza, M. (2013). Does auditor industry specialization improve audit quality? *Journal of Accounting Research, 51*, 779–817.

Myers, J., Myers, L., & Omer, T. (2003). Exploring the term of the auditor-client relationship and the quality of earnings: A case for mandatory auditor rotation? *The Accounting Review, 78*(3), 779–799.

Reichelt, K. J., & Wang, D. (2010). National and office-specific measures of auditor industry expertise and effects on audit quality. *Journal of Accounting Research, 48*(3), 647–686.

Singer, Z., & Zhang, J. (2018). Auditor tenure and the timeliness of misstatement discovery. *The Accounting Review, 93*(2), 315–338.