Global Information Assurance Certification Paper

Defense in Depth: Can Geolocation Help Prevent Tax Fraud?

GIAC (GSEC) Gold Certification

Author: Jon Glas, jrg644ss@yahoo.com
Advisor: *Johannes Ullrich*
Accepted: *November 25, 2019*

Abstract

Accountants and tax filing businesses use complex software to automate the preparation and electronic filing of tax returns. Cybercriminals harvest identities, breach networks, and impersonate legitimate users to leverage tax software to defraud the government, the affected businesses, and citizens for over $1 billion annually (McTigue, 2018). The IRS and tax software companies have partnered to implement controls focused on authentication, authorization, and detection to identify fraudulent tax returns before they are processed. These controls successfully prevent upwards of $10 billion of fraudulent filing a year (McTigue, 2018), but those controls focus on an analysis of the 'who' and 'what' components of tax returns. This paper uses Geolocation tools to look at the 'where' component of tax returns by analyzing legitimate and fraudulent tax return electronic filing data to look for trends and patterns. The goal of this paper is to determine if Geolocation technologies can be used as an additional layer of controls to support a defense in depth approach of fraud prevention.

1. Introduction

Benjamin Franklin said: “Our new Constitution is now established, and has an appearance that promises permanency; but in this world nothing can be said to be certain, except death and taxes” (Smyth, 1907, p.69). It is indeed true that taxes are a certainty for Americans. The requirement to pay taxes has created industries that facilitate calculating and filing taxes for individuals and businesses.

Each American must file their taxes annually with the Internal Revenue Service (IRS) of the U.S. government based on certain requirements, including marital status and gross income (Erb, 2019). The tax code is complex. The amount of taxes individuals are required to pay varies based on numerous factors. Paper tax forms are filled out by hand and mailed to the IRS for processing. Paper filings processing can take weeks, or even months, to complete. The complexity and volume of tax filings that require processing have paved the way for innovations in the automation of tax calculations and the filing of taxes digitally. Companies, including H&R Block, TaxAct, Intuit, and Drake, have built software solutions to solve this need for automation.

Accountants and other tax filing businesses use these software solutions to create and automate the electronic filing of tax returns, also known as e-filing. These computer systems greatly simplify the effort of filing taxes. Once a user enters their financial data into the system, online tax forms are completed and sent to the IRS automatically. The IRS then reviews digital filings and issues tax refunds for overpaid taxes.

These systems have streamlined the work required to file taxes. However, they have also drawn the attention of criminals seeking financial gain. These criminals exploit a critical flaw in the IRS’ system. The IRS issues tax refund payments based on the data entered into the form. By entering fraudulent information, a user can trigger an IRS refund illegitimately. While the IRS must accept a tax filing before issuing payment, there are too many filings for it to process each filing thoroughly. Instead, the IRS only performs a cursory validation of returns to determine whether to accept them.

In order to extract payments, cybercriminals must pass the initial IRS checks. Refunds require that someone has already paid taxes to the government. Creating

fraudulent data from scratch is not viable, so cybercriminals prey upon individuals and businesses. Cybercriminals harvest identities, breach networks, and impersonate their victims to create fraudulent tax return filings.

Detecting and preventing this type of fraud is difficult. Both the private sector and the IRS have implemented several techniques to address this threat. These techniques include both identity verification tools, such as two-factor authentication and anomaly detection using the Return Review Program. This program relies on preset rules and pattern recognition to flag questionable returns. Nevertheless, despite these controls, the IRS is defrauded for over $1 billion annually (McTigue, 2018).

The IRS and the private sector have implemented a Defense in Depth-methodology in the Revenue Return Program for combating this growing problem of tax return fraud (McTigue, 2018). Authentication techniques already strive to verify the 'who' component of tax returns, which verifies the person submitting a return is who they claim to be. Analysis techniques such as anomaly detection aim to verify the 'what' component of tax returns, which checks that what is in the tax return is not out of the ordinary. These approaches are insufficient to combat fraud, and new tactics, tools, and techniques are required. One component that is lacking is the 'where' component of tax returns, which verifies the location used in a tax return filing is not suspicious. A technology that can help in this regard is Geolocation. Geolocation uses either GPS or location data stored for IPs to identify where network traffic originates. It also can be leveraged to determine where e-filings occur. The question is whether this data can be analyzed and leveraged to detect fraudulent tax return filings?

This research paper evaluates modern Geolocation technologies, their accuracy, and reliability. It will analyze tax filing data to explore functional ways of using Geolocation technologies to detect fraud. The research will conclude the viability of using Geolocation technologies as a defense mechanism against tax fraud.

2. Geolocation Overview

Geolocation is the process or technique of using the digital information of a person or device to identify a physical location (Muir & Oorschot, 2009). Geolocation techniques determine location data using GPS, triangulation, network analysis, and web APIs to determine location data.

Jon Glas, jrg644ss@yahoo.com

The first uses Global Positioning Systems (GPS) to identify the physical latitude and longitude of a location. Cellular phones, laptops, and tablets come equipped with GPS tools that can identify the coordinates of the physical location of the device.

Wireless and cellular triangulation is another technique used in Geolocation. Wi-Fi Positioning System (WPS) can approximate a location based on metrics of the wireless signal itself. Cellular devices using wireless or cellular signals will communicate with multiple cellular signal towers within reach of the device's wireless signal. The physical location can be estimated based on an analysis of the signal data between the towers (Abdou & Van Oorschot, 2018).

Several techniques can physically locate a Public IP address. Public registries such as the WHOIS database or private providers, including MaxMind, Neustar IP Intelligence, and IP2Location, have services that provide physical location data for a given IP address (Abdou & Van Oorschot, 2018). Dynamic Geolocation can be performed against Public IP addresses as well. Similar to Wi-Fi signal analysis, network metrics can be analyzed to provide metadata used to estimate a physical location. Network data such as routers passed, transmission times, and network headers can be analyzed to estimate a physical location.

Most web browsers fully support Geolocation API functionality. Websites make calls through the built-in HTML5 Geolocation APIs to get the latitude and longitude of the user's position. This supported Geolocation API is optional and requires user permission to function (Abdou, A., & Van Oorschot, P. C., 2018).

2.1. Tax Fraud Overview

Tax fraud involves misleading the IRS into issuing a tax refund using the tax filing process. Filers must show that an individual paid the IRS a higher sum in taxes than they are required to receive a refund. Fraud often involves identity theft, referred to as SIRF, or stolen identity refund fraud (C. Denton, personal communication, September 14, 2019). Criminals target both individuals and corporations for identity theft. The substantial quarterly tax payments of large corporations make them particularly attractive targets for SIRF. When filing their corporate taxes, companies may receive a multimillion-dollar refund. Criminals committing SIRF will file these

taxes to maximize the refunds and reroute the payment to accounts they own (C. Denton, personal communication, September 14, 2019).

2.1.1. Types of Fraud

This research focuses on the three types of SIRF: individual account takeovers, professional account takeovers, and created accounts (C. Denton, personal communication, September 14, 2019). Personal fraud does not qualify as SIRF even though it involves inaccurately filing taxes and is beyond the scope of this research.

1. **Individual account takeovers** involve cybercriminals compromising a personal account in a tax software system. The criminals change payment routing data in the tax return. They then either submit the return or wait for the account owner to submit it themselves.
2. **Professional account takeovers** involve hackers compromising an account in a professional tax software suite used by a business or independent accountant. After gaining access to all client identities, the hackers search for unfiled taxes and alter refund routing numbers.
3. **Created accounts** involve criminals acquiring the tax software themselves and creating an account using a stolen identity and consumer licenses of tax products. Hackers can use network breaches to steal datasets for professional software suites. These require professional software systems and accounts with purchased entitlements, which are used by the software to track permitted filings. Stolen datasets can be imported and used in fraudulent filings. Criminals use stolen credit cards to pay for software and licensing costs, causing chargebacks against the software company.

2.1.2. Types of criminals

Along with the different types of fraud, there are different types of criminals who commit SIRF. The most common types of SIRF criminals are tax professionals, business insiders, and remote hackers.

Unethical or criminal tax professionals commit fraud through their businesses and accountancies. They lie to their clients and keep a portion of their clients' refunds for themselves. Occasionally, cybercriminals create temporary businesses to file taxes for clients at attractive prices. These businesses operate for several weeks before

disappearing overnight, taking services fees and refunds with them. (C. Denton, personal communication, September 14, 2019)

Malicious insiders use the legitimate businesses they are employed in to commit fraud. These insiders either have elevated access to the business's accounts or use harvested credentials to gain access and commit account takeover fraud (C. Denton, personal communication, September 14, 2019).

Remote hackers are cybercriminals who operate remotely and use modern hacking techniques to gain access to a personal account or breach a business's network. Criminals harvest credentials and steal datasets for use in SIRF (C. Denton, personal communication, September 14, 2019).

2.1.3. Fraud drivers

The reasons for fraud fall into two categories: financial and sabotage. Financially incentivized fraud, where the criminal is seeking financial gain, makes up the vast majority of fraud cases. Unfortunate circumstances, desperation, and greed typically drive financial fraud. Sabotage is where a malicious insider or remote hacker seeks to hurt a business or an individual. Revenge and hate motivate sabotage, which makes up a small portion of fraud cases (C. Denton, personal communication, September 14, 2019).

2.1.4. Fraud payments

Tax fraud is a viable means of financial gain due to the payment options available to consumers. The IRS uses standard SWIFT payment networks that only require routing and account numbers allowing criminals to set up reloadable, anonymous debit cards. Owners of the cards liquidate payments quickly to avoid consequences of fraud (C. Denton, personal communication, September 14, 2019).

2.1.5. Existing fraud controls

The IRS and the private sector have already implemented many controls to detect and prevent fraud. These controls fall into two categories: mandatory and proactive.

Mandatory controls are regulatory laws and compliance frameworks required by tax agencies such as the IRS or state Departments of Revenue. These controls (C. Denton, personal communication, September 14, 2019) include but are not limited to:

- Username and password complexity and reset policies
- Multifactor authentication requirements
- Forced login and idle timeouts
- Social Security Number reuse & EFIN verification checks
- Product and schema metadata tracking used to produce a risk score
- NIST framework compliance, required by all states.
- DISA STIGS compliance required by the Indiana Department of Revenue. STIGS are Secure Technical Implementation Guides that list required system configurations developed by the Department of Defense to secure critical systems.
- IRS reporting requirements. Tax companies cannot legally identify fraud and must instead submit reports of fraud leads to the IRS, the final arbitrator of fraud.

In addition to the mandatory controls required by federal or state governments, many companies implement additional controls to combat fraud. Some of these controls build on mandatory controls (C. Denton, personal communication, September 14, 2019). These include:

- The metadata tracking controls used in the product and data schema are enhanced to include additional checks, such as behavioral screens to look for anomalies in expected behavior.
- Threat intelligence is also heavily leveraged to keep bad actors out of the system by maintaining blacklists of IPs and accounts and using analytics to track aliases and usage patterns.
- Account recovery processes attempt to leverage more secure methods to verify users, such as requiring driver's licenses or other forms of photo identity verification.
- Behavioral pattern analysis attempts to block fraudulent usage. Patterns include blocking or closely scrutinizing questionable sales, such as large purchases towards the end of the tax season (C. Denton, personal communication, September 14, 2019).

Jon Glas, jrg644ss@yahoo.com

3. Research Method

Understanding the potential impact Geolocation can have on tax software fraud requires examining the advantages and disadvantages of the technology to see how it applies to real-world scenarios. Two years of customer tax filing data were collected for analysis from a prominent software company in the tax filing industry. This company's name and its data have been redacted and scrubbed to protect its identity.

The data collected is from the 2017 and 2018 tax years. This data includes a total of 91,789,238 filing records spanning both tax years. The data elements collected are limited to fields required to perform Geolocation analysis. Sensitive customer data, such as name or social security number, are not included to protect the identities of the filers and the firms that filed their taxes.

Table 1 below and Appendix A contain the data fields and types used in this research. The researcher imports scrubbed filing records extracted from the source company into a MySQL database table called Data. The scrubbing process replaced the existing Business Identification Number (BIN) value with a randomized replacement value to keep client data anonymous. Table 1 contains the data structure of the fields contained in the Data table.

Field	Notes	Type	Length	Key
FILING_ID		BIGINT	11	PRI
BIN	Account number	VARCHAR	255	
CLIENT_IP_ADDR		VARCHAR	255	Foreign
RX_CLIENT_DATE		DATETIME		
IP_TWO_LETTER_COUNTRY		VARCHAR	255	
PAYMENT_AMOUNT		INT	11	
FILING_STATE	Accept/Reject?	INT	11	
IP_CREATE		VARCHAR	255	Foreign
IP_SUBMIT		VARCHAR	255	Foreign
FILING_TYPE		VARCHAR	255	

FRAUD		VARCHAR	255	

Table 1 - Data Table Structure

The data provided did not contain any GPS data. Instead, it contained three IP address fields: CLIENT_IP_ADDR, IP_SUBMIT, IP_CREATE. These IPs were extracted and filtered for uniqueness and subsequently ran through a Geolocation lookup service to generate the Geolocation data for each IP. The service is a wrapper to the MaxMind Geolocation service that keeps a database mapping IPs to physical locations. The resulting records were imported to a separate table called Ipgeo. Appendix A contains the data structure for the Ipgeo table.

The intended analysis focused on fraudulent transactions and the clients that performed them. Because queries against the 91 million records in the Data table were lengthy, the records searched needed to be filtered down. The Data table contained the Fraud field, denoting if a filing record was fraudulent. The researcher identified a list of BINs containing fraud. All filing transactions for those BINs were extracted and imported into a separate table called Evildata, which duplicate the structure of the Data table. The Evildata table contained both fraudulent and legitimate transactions for the selected BINs but omitted records from BINs that did not have any records flagged as fraudulent. The Evildata table was then indexed based on the fields containing IP addresses, and a foreign key relationship established to the Ipgeo table's IPAddress primary key field.

Several queries were run against the Evildata, Data, and Ipgeo tables using Join clauses to combine the tables. The query execution was slow, and the Evildata table was combined with the Ipgeo table into the Evildata_clientip_geo table to save time on Join queries.

The aim of analyzing this data was to identify trends and patterns in how clients were filing data and how fraudulent transactions fit into these trends. The data points discovered during this analysis will provide insight into new questions or patterns that could identify controls for detecting and preventing fraud. Statistical analysis performed determines whether a correlation between the tax fraud and geolocation data for each of the firms exists. Conversely, the analysis could also reveal roadblocks in further analysis of a trend or datapoint.

Jon Glas, jrg644ss@yahoo.com

4. Findings and Discussion (Exposition of the Data)

4.1. Baseline fraud findings

The first step was to generate baseline summary data derived from the total dataset. There were 96,802 total BINs in the two years' worth of filing data, of which 3,896 BINs, 3.99%, contained fraudulent transactions. The total dataset contained 91,789,238 total records, of which 9,068 identified as fraudulent. The dataset shows that while approximately 4% of the total entities that file using the company's software involved some form of fraud, only 0.00988% of those total transactions were fraudulent.

The filing refunds totaled $112,640,712,256. The fraudulent filing refund totaled $35,093,991, only 0.03115% of the total. The highest fraudulent refund was $543,716, with the lowest fraud amount being a payment, not a refund, of $194,757. The listed high and low gives an average fraudulent refund total of $3,822.02. The highest legitimate refund amount was $69,600,976, with the highest payment amounting to $76,567,958. The average refund amount of all legitimate filings was $104.94.

$35 million is a sizable fraudulent sum over two years. However, this amounts to a very small percentage of the whole. This amount is spread across all the fraud types detailed in Section 2.2, which limits the available scope for this research.

Geolocation fraud controls are only effective in identifying fraudulent filings based on location. Fraudulent professionals are malicious account owners who have permissions to whitelist locations or ignore warnings raised by any Geolocation fraud controls implemented. Malicious insiders may also have acquired these necessary permissions. If insiders have access to the account owner's place of business, they could commit their fraudulent transactions on-premise and bypass any proposed Geolocation controls.

A small subset of malicious insiders who commit fraud could choose to file fraudulent tax returns outside of normal filing locations. These insiders would mirror the behavior of remote attackers and would be grouped in with them. This type of fraud, account takeover by a remote attacker, will be the focus of further research.

Jon Glas, jrg644ss@yahoo.com

4.1.1. Data integrity concerns

It is important to mention two main concerns regarding fraud data. Transactions flagged as fraudulent may contain little data indicating the reason why it is fraudulent. When the IRS reports an e-file submission for an identity that has already had a return filed for it that tax year, it marks both filings as fraudulent. If the filings were both done using the same tax software, two fraud records would be listed in the analysis data when only one is fraudulent. This double reporting is impossible to identify in the analysis data as identity data is not present. The search, therefore, assumes that all listed fraud data is actual fraud for analytical purposes.

The second concern is the focus on IP for Geolocation data. The analysis data contains two years of stored transaction data. The IPs stored for those filings could have changed owners and locations. The Geolocation lookup via MaxMind's database was performed in September 2019 and returned updated location data for those IPs. Dated lookups are a concern when analyzing IPs used by the same entity and the corresponding Geolocation data, but it is not a blocker for analysis. The purpose of this research is to identify if Geolocation could help identify possible fraud. The data may be skewed based on the dated IP Geolocation lookups, but the real-world data still provides valid situations that could arise in the industry and can provide value.

4.2. IP Analysis

The analysis identified 4,386 IPs used in fraudulent filings. The filings are broken down into the number of fraudulent filings committed by each IP.

Fraudulent filings per IP	IP count
Less than 9	4276
9 to 24	69
25 to 99	29
100 to 199	2

Table 2 - Fraudulent filings per IP

Most IPs commit less than nine fraudulent filings per IP. The higher number of fraudulent filings could be a factor in showing either massive account compromise or entirely fraudulent BINs. This research will compare the fraudulent transactions against legitimate transactions to better understand these counts.

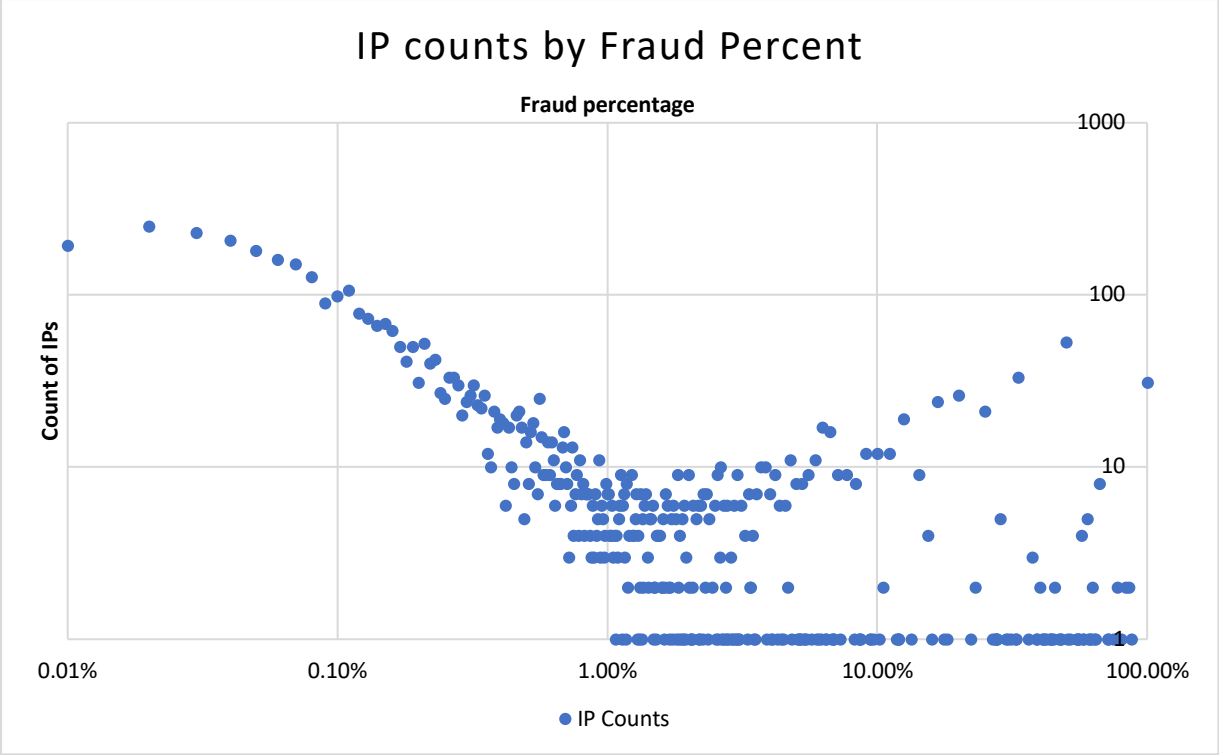

Figure 1 – IP Counts by Percentage of Fraud – Logarithmic

Figure 1 shows how many IPs have the same percentage of fraudulent filings. The graph shows a weighting to the left, where 70.57% of all IPs have a fraud rate of 10% or lower, while 52.32% of IPs have a fraud rate below 5%. The graph shows that a large majority of the fraudulent filings are a low number of occurrences and, therefore, those IPs have a high number of legitimate filings. IPs with both fraud and non-fraud filings are an important distinction: if an IP has a large percentage of legitimate filings, the fraud Geolocation data would be identical to their non-fraud Geolocation data. Considering that most of the transactions on a low fraud IP are legitimate rules out the option of identifying a fraudulent transaction through whitelisting approved locations and making proximity comparisons to a whitelisted location. The IP of the fraudulent and legitimate filings would be identical. Finding patterns to identify fraud would require looking at other non-geo data, effectively eliminating these as cases for this research.

This analysis shows an inverse relationship between the percentage of fraud and how effective Geolocation is at detecting that fraud. Looking at the right side of Figure 1 shows 31 IPs with 100% fraud. The next IP is 87.5%, followed by two IPs with 85.71%. The pattern continues with only one or two IPs at various percentages until 66.67%, where there are eight counts of fraud followed by a 53 IP spike at 50%

fraud. Only 0.84% of IPs have over 80% fraudulent transactions, and only 0.87% of IPs have between 50% and 80% fraudulent transactions.

This distribution allows the breakdown of fraud distribution into three groups: high, medium, and a low percentage of fraud. The low percentage group is not in scope for further analysis as the geodata of fraudulent data is identical to non-fraudulent data. The medium percentage group of IPs between 10% and 50% fraud forms a grouping that shows significant fraud but still contains legitimate filings and shows the same geodata issues exist as with the low percentage group. These could include any of the fraud types. The medium percentage IPs need further analysis to determine if their geolocation data can separate fraud from the normal transactions of the business. The high percentage group of IPs with fraud higher than 50% is where Geolocation analysis can prove the most valuable.

4.3. BIN analysis

Trend analysis continues by looking at the number of fraudulent filings by BIN. A total of 3,868 BINs contain fraud, which is further broken down into groups by BIN.

Fraud filings per BIN	Count
Less than 9	3747
9 to 24	89
25 to 99	27
100 to 199	4
200+	1

Table 3 - Fraudulent Filings per BIN

BINs show a similar trend to IPs. 96.87% of occasionally fraudulent BINs commit eight or fewer counts of fraud each. Sorting the BINs by percent of fraud shows that only 30 BINs have fraud transactions of 10% or higher. We extract those top 30 BINs and evaluate the fraud breakdown by IP, as shown in Figure 2.

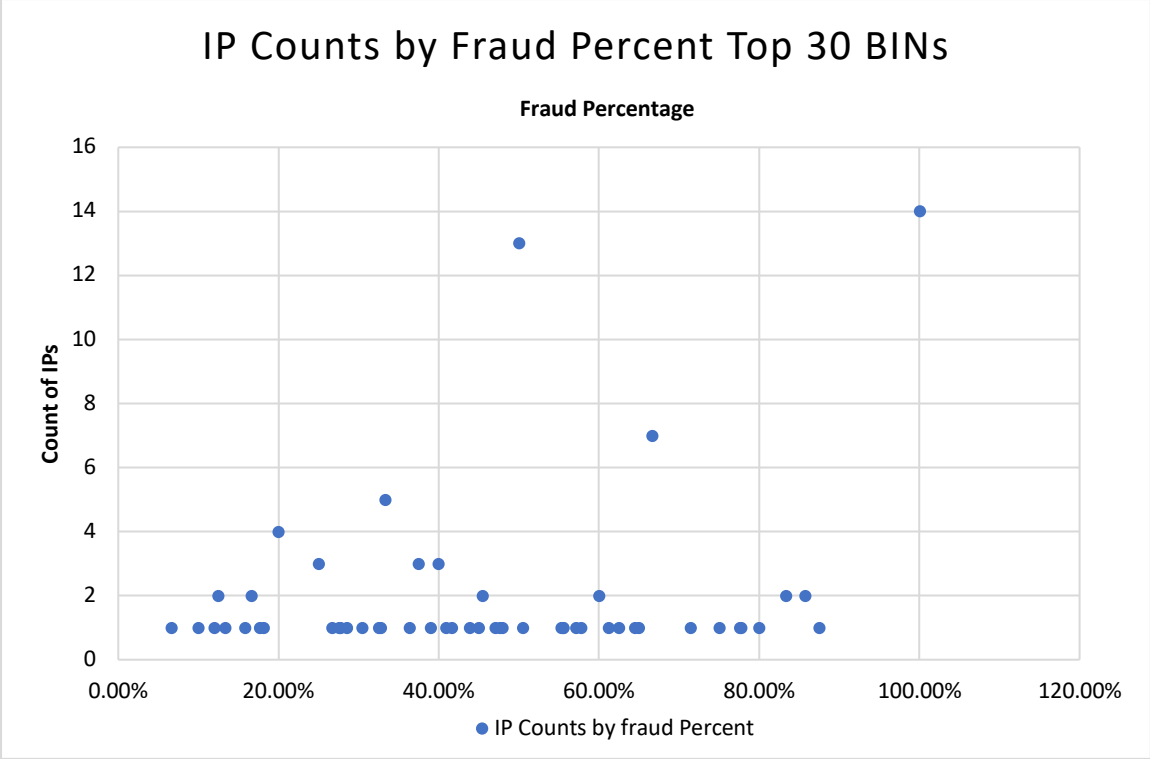

Figure 2 - IP Counts by Fraud Percentage for Top 30 BINs

The top 30 BINs have a more even distribution of fraud percentages, with 56 IPs falling in the 50% or greater group and 46 IPs falling in the 10% to 50% group. Only a single IP falls in the small group of less than 10% fraud. This sample of BINs provides a good starting point to focus on for geolocation analysis.

The top 30 BINs committed 90% of the fraudulent transactions. There are still IPs with high fraud percentages that do not fall in the top 30 fraudulent BINs. 81 IPs fall in the high fraud group of greater than 50% fraud and 174 IPs that fall in the medium fraud group of 10% to 50% fraud. These counts are not accounted for by the top 30 fraud BINs and require further evaluation.

4.4. Evaluation of TOR

Geolocation data from MaxMind contains anonymous data, such as TOR exit node IPs. As TOR is not a typical use case for tax professionals using tax software, any TOR exit node filings would be suspect (C. Denton, personal communication, September 14, 2019). In the two years of tax data, five total IPs are TOR exit nodes, but none of the filings associated with those IPs were fraudulent. TOR exit nodes change frequently. While these IPs were TOR exit notes at the time of the Geolocation lookup in September of 2019, it appears likely that they were not in 2017

and 2018. No further analysis is necessary regarding the five IPs flagged as TOR, but the TOR exit node metadata is a useful data point for calculating the risk scoring of transactions.

4.5. Geolocation Analysis

4.5.1. Baseline geolocation data

The primary hypothesis is to determine if there is a difference between fraud geolocation data and non-fraud location data that shows a pattern that could flag future filings as suspect. The latitude and longitude geolocation data were queried for each IP and grouped by the BIN. The researcher scanned Each BIN to count how many of their associated IPs were from different coordinates, which equated to 2753 total BINs, which is 71.16% of total BINs containing fraud. Twenty-four of the top 30 fraudulent BINs had IPs with different coordinates. The remaining six BINs were confirmed to only include a single latitude/longitude and thus could be eliminated from further analysis.

The analysis included latitude and longitude comparisons but also looked at the top 30 fraudulent BINs to determine the BIN counts per IP. There were 131 IPs from the top 30 BINs that were each found to include filings from multiple BINs. Ten IPs each had over 100 different BINs with one IP having 428 different BINs. Shared IPs complicate the analysis by showing there are cases where an IP has a one-to-many relationship with clients.

Querying the data was proving to be time-consuming, so the fraud data was isolated into temporary tables, which accelerated analysis. A query of the overall Evildata table to pull all IPs and filing counts grouped by the IP, BIN, and fraud fields yielded 52,219 rows. This data was filtered down to the fraudulent BINs that were in the top 30 while containing multiple latitude and longitudes in its geolocation data. The rows were filtered down to 249 rows of IP/BIN combinations. Below is a sample of the data.

IP	BIN	count(*)	fraud	In Top 24 w multi-lat/long
107.158.12.198	1	2	NULL	y
107.174.133.170	2	40	NULL	y
107.174.133.170	2	20	Fraud	y

107.77.197.91	3	1	NULL	y
107.77.197.91	3	2	Fraud	y
107.77.199.91	3	2	NULL	y
107.77.199.91	3	1	Fraud	y
107.77.215.112	4	13	NULL	y
107.77.215.112	4	2	Fraud	y
107.77.215.140	4	2	NULL	y
107.77.215.142	4	13	NULL	y
107.77.215.142	4	5	Fraud	y
107.77.215.166	4	5	NULL	y
107.77.215.166	4	5	Fraud	y
107.77.215.171	4	14	NULL	y
107.77.215.171	4	2	Fraud	y
107.77.215.175	4	2	NULL	y
107.77.215.226	4	1	NULL	y
107.77.215.43	4	1	NULL	y
107.77.235.19	4	1	NULL	y

Table 4 - IP in the top 24 fraud BINs with multiple Latitudes/Longitudes

The extracted data contained 153 unique IPs, and a query of those IPs showed 3717 filings from the 24 BINs top fraudulent BINs. The same query ran against BINs that contained fraud returned 3775 filings. The difference in results indicated that the IPs from the top 24 fraudulent BINs also completed filings in BINs that were not part of the top 30 fraudulent BINs.

Analysis of the filings from these 153 IPs looked for patterns related to fraudulent and non-fraudulent transactions. IP comparisons performed between the variations between the CLIENT_IP_ADDR, IP_SUBMIT, and IP_CREATE fields. IP_CREATE was the IP of the client used to create the filing record. The IP_SUBMIT was the IP used to submit the filing record to the IRS. The CLIENT_IP_ADDR matches the IP_SUBMIT value in all cases where the IP_SUBMIT field is present in a record, which is not always the case. When looking at the fraudulent filings, 30.45% of the filings had a CLIENT_IP_ADDR value that was different than the IP_CREATE field. While this is a significant percentage, it is not a strong correlation to differentiate a fraud filing from a non-fraud one.

Next, the analysis looked at geolocation data from the top 24 BINs that contained multiple filing locations. In all cases, including the highest fraud percentage IPs, all occurrences of latitude and longitude values for the IPs were present in both the fraudulent and legitimate filing records. A relevant example is BIN 2279, which

contains 174 fraudulent filings out of 1071. Six IPs had different latitudes and longitudes in the filings from 4 different cities in Arizona: including Phoenix, Mesa, Scottsdale, and Tucson. Four hundred twenty-eight of the filings were from Arlington Heights, Illinois. This data point is notable because it's geographically distant from the other IP locations. However, there were zero fraudulent transactions filed from this location, which hindered attempts to identifying fraud patterns. The other four geolocations were in relative proximity to each other in the state of Arizona. The fraud transactions were spread out between the IPs along with legitimate filings.

The number of legitimate transactions mixed in with the fraudulent filings under the same IP makes identifying a pattern impossible with the data in this study. All the BINs checked for geolocation patterns follow this trend, showing that compromised accounts from the BIN owner's network are prevalent. This analysis eliminates Geolocation as an identifying factor, given the data available for analysis.

4.5.2. Statistical Analysis of Geolocation Fraud

Pattern recognition can be aided by statistical analysis of each firm in our data sample to determine if there is a correlation. The firm fraud percentage is analyzed against the mean distance between the geolocations identified for the firm's IPs. The geolocation data contains latitude and longitude coordinates. The Haversine formula below calculates the distance between two points over the earth's surface (Veness, 2012):

Haversine formula:

$$a = \sin^2(\Delta\varphi/2) + \cos\varphi_1 \cdot \cos\varphi_2 \cdot \sin^2(\Delta\lambda/2)$$

$$c = 2 \cdot \operatorname{atan2}(\sqrt{a}, \sqrt{(1-a)})$$

$$d = R \cdot c$$

where φ *is latitude,* λ *is longitude,* R *is earth's radius (mean radius = 6,371km);*

Figure 3 - Haversine Formula

A python script, shown in Appendix B, applies this formula to calculate the distances between all the geolocation coordinates for each BIN in the dataset. It then generates the mean of those distances. The BIN's fraud percentage is determined through a SQL query shown in Appendix C. These values were used to calculate a Correlation Coefficient that measures a linear dependence between the two variables using the Pearson Correlation Coefficient Formula detailed in Appendix D (Correlation Coefficient Calculator, 2019). The data sample contained 2753 BINs and

resulted in a Correlation Coefficient of -0.017215. Results close to -1 represent a strong negative correlation, and results close to 1 represent a strong positive correlation. The result of -0.017215 is very close to 0, which represents no correlation. The graph below shows the correlation results in more detail:

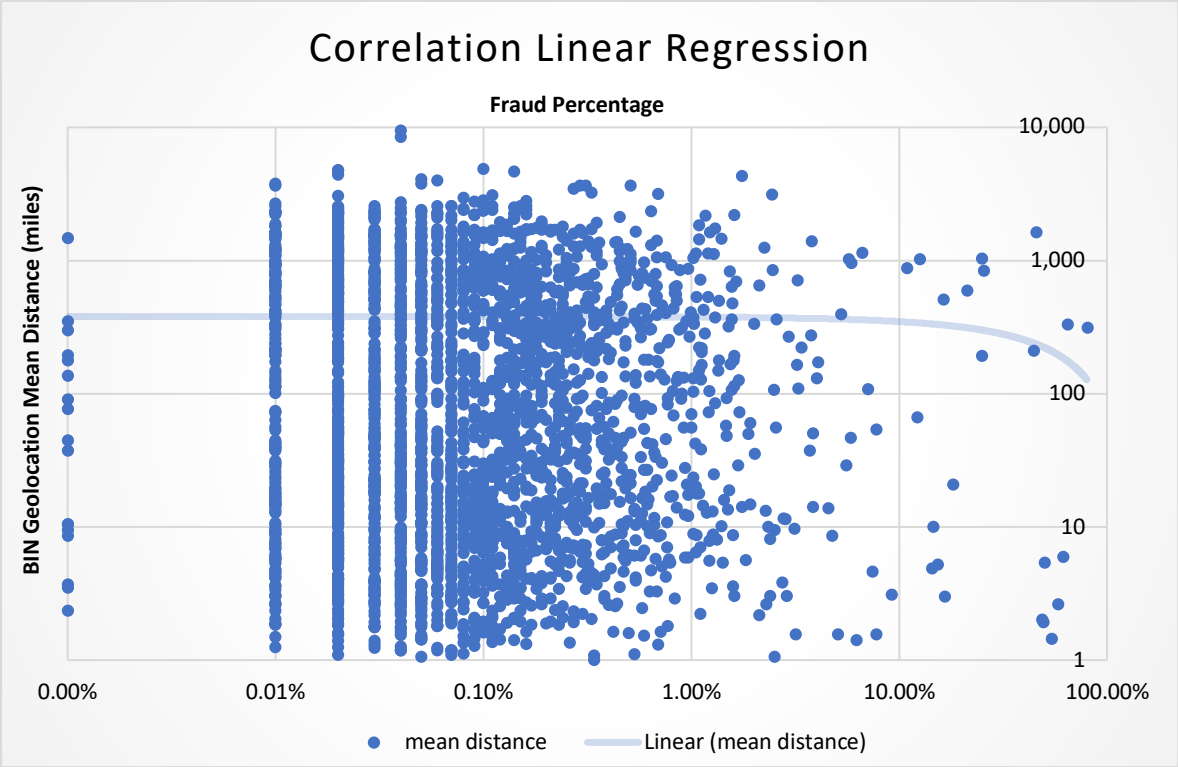

Figure 4 - Correlation Linear Regression – Logarithmic

The vertical axis in Figure 4 shows the mean distance, while the horizontal axis details the fraud percentage. Each plot point is a single BIN. The graph is heavily weighted, showing most of the fraud percentages below 10%, which corresponds with our IP and BIN analysis. The plot points for distances show an even distribution. The linear regression line is flat, dropping as it reaches the higher fraud percentages on the right of the graph. The flat regression line supports the Correlation Coefficient in showing there is not a linear correlation between the fraud percentage of a BIN and the distance between that BINs IP geolocations. This lack of correlation supports the improbability of distinguishing between fraudulent and legitimate tax filings.

4.6. Industry Resistance

The data collection practices of Google and Facebook have raised significant concern from the end-users of the software. These concerns are very serious in the tax industry, where highly sensitive and critical data is required. Because of this, there is

resistance to adding new features that require data collection to tax software by (C. Denton, personal communication, September 14, 2019). Implementing a feature that captures geolocation data in a way that appears invasive to users is rejected even if those features are optional or promote security overall.

This stigma against data collection makes optional controls like geolocation lookup through the HTML 5 browser support unreliable as users are resistant to give it permissions to look up their location. Crafting more robust solutions for geolocation lookup that use GPS does not have an acceptable return on investment as the adoption of those solutions would be limited.

The restrictions imposed by the industry requires geolocation to use IP lookup through IP location database providers. The accuracy of this method of geolocation lookup has been questioned for the past ten years. Initial studies have shown that IP databases were accurate at the country level, but more granular lookups were inaccurate (Poese, Uhlig, Donnet, Kaafar, & Gueye, 2011). This accuracy has improved over the past eight years. Verification of a sample of 25 clients from the source data determined that IP lookups through the MaxMind service were 100% accurate down to the city level. However, none of the lookups had the correct street address, latitude, or longitude of the client's listed location. Five of the lookups were within four miles of the client's physical address, with the remaining 20 being within 10 miles. The resulting geolocation data look to be a central hub of the internet service provider rather than the physical address of the client.

Evasion techniques by attackers are also a major concern in geolocation reliability. Location data is subject to manipulation using browser extensions such as Firefox's Fake Location and Location Guard, which forge the location data from the browser before being passed onto any requested server (Abdou & Van Oorschot, 2018). If a simple IP lookup is used, which is prevalent in the tax industry, then the actual device IPs of the user can be masked behind a proxy or some other form of anonymizer. When you introduce evasion techniques and manipulation by an adversary, then geolocation service success rates will be low (Muir & Oorschot, 2009).

The Client Presence Verification technique uses servers to verify geolocation falls within a triangulated area between three of the servers, called verifiers (Abdou &

Van Oorschot, 2018). This technique does have the potential to thwart forged geolocation data. However, this technique requires the verification servers to communicate with the source to measure metadata about the connection to determine if they are indeed within an expected area. This type of verification would experience heavy resistance from end-users of tax software for the aforementioned data collection concerns. Cost is also a significant detractor for this CPV technique as potential clients span the entire country, and hosting the required verifiers to be able to cover the entire country is cost-prohibitive. This technique also doesn't solve false negative issues that would occur if a remote hacker utilized a proxy in the triangulated area or filed their transactions from the client network and location.

4.7. Recommendations for Practice

Geolocation is not mature enough to warrant its use as a primary detection tool for tax fraud. The resistance from the customer base to having their information captured is a large detractor from using the technology. This resistance forces feature implementation into a narrow list of options like IP location databases, which are inaccurate, dated, and easily bypassed by a malicious actor. The technology needs to be hardened and adopted in the culture before it can be a true asset.

SIRF is problematic for crafting patterns. Fraud quantity is very low and uses the same IPs as legitimate traffic, making it impractical to identify using geolocation data. The best use of geolocation at this time is as a component of a larger risk scoring system. Layering it alongside other risk factors can help calculate risk but doesn't rely entirely upon geolocation as a singular source of detection.

4.8. Implications for Future Research

Dated research is a concern for Geolocation data. Geolocation data needs to be stored along with other filing data at the time of the transaction as IP databases change over time, and future geolocation lookups may result in erroneous data.

Tax fraud detection is a prediction problem. Artificial intelligence and machine learning are the most suitable technologies for predictions. As more data is stored and fed into AI/ML models, the models become more accurate and faster than manual data analysis. The challenges encountered in this study conclude that further research should utilize AI/ML technologies.

Jon Glas, jrg644ss@yahoo.com

5. Conclusion

Tax fraud is a billion-dollar problem that the public and private sectors must deal with on an annual basis. Many controls are already in place to try to detect and prevent stolen identities and identify tax fraud. Geolocation is one technology that is not in heavy use in the effort to identify tax fraud. The evaluation showed the industry itself to be heavily resistant against the collection of sensitive data such as location data. This resistance limits the implementation methods for Geolocation, such as using IP database lookups versus GPS. The accuracy of IP databases is lacking and only useful down to the city level. Additionally, network changes can render IP databases outdated. Geolocation is easily bypassed by malicious actors using readily available tools and proxies to mask their true location.

Geolocation isn't mature enough to be a primary detection mechanism in the fight against tax fraud. Geolocation is an informative tool that can help with risk. It is not reliable enough to warrant investment as a detection tool in its current state.

Jon Glas, jrg644ss@yahoo.com

References

Abdou, A., & Van Oorschot, P. C. (2018). Secure Client and Server Geolocation over the Internet. *;login:, 43*(1), 19-25. Retrieved from https://www.usenix.org/publications/login/spring2018/abdou

Cahill, M.H., Lambert, D., Pinheiro, J.C., & Sun, D.X. (2002). Detecting fraud in the real world. In *Handbook of massive data sets* (pp. 911-929). Springer, Boston, MA

Correlation Coefficient Calculator (2019) Retrieved from http://www.endmemo.com/statistics/cc.php

Dyck, W. (2012, June 6). *Haversine.py*. Retrieved from https://gist.github.com/rochacbruno/2883505

Erb, K. P. (2019, January 30). Do You Need To File A Tax Return In 2019?. Retrieved from https://www.forbes.com/sites/kellyphillipserb/2019/01/30/do-you-need-to-file-a-tax-return-in-2019.

McTigue, J. R. (2018, August 23). Tax Fraud and Noncompliance: IRS Could Further Leverage the Return Review Program to Strengthen Tax Enforcement. Retrieved from https://www.gao.gov/products/GAO-18-544.

Muir, J. A., & Oorschot, P. C. (2009). Internet geolocation. *ACM Computing Surveys, 42*(1), 1-23. doi: 10.1145/1592451.1592455

Poese, I., Uhlig, S., Kaafar, M. A., Donnet, B., & Gueye, B. (2011). IP geolocation databases: Unreliable? *ACM SIGCOMM Computer Communication Review, 41*(2), 53-56. doi: 10.1145/1971162.1971171

Smyth, A. H. (1907). *The writings of Benjamin Franklin* (Vol. X). New York: Macmillan

Jon Glas, jrg644ss@yahoo.com

Veness, C. (2012). *Calculate distance, bearing and more Latitude/Longitude points.* Retrieved from https://www.movable-type.co.uk/scripts/latlong.html

Jon Glas, jrg644ss@yahoo.com

Appendix A

Table ipgeo data structure

Field	Type	Length	Key
IPAddress	VARCHAR	255	PRI
lat	VARCHAR	255	
lng	VARCHAR	255	
alt	VARCHAR	255	
houseNumber	VARCHAR	255	
street	INT	11	
city	INT	11	
administrativeAreaLevel3	VARCHAR	255	
administrativeAreaLevel2	VARCHAR	255	
administrativeAreaLevel1	VARCHAR	255	
administrativeAreaLevel1Short	VARCHAR	255	
country	VARCHAR	255	
countryShort	VARCHAR	255	
postalCode	VARCHAR	255	
companyName	VARCHAR	255	
cityConfidence	VARCHAR	255	
region	VARCHAR	255	
regionCode	VARCHAR	255	
isCountryEmbargoed	VARCHAR	255	
regionConfidence	VARCHAR	255	
countryConfidence	VARCHAR	255	
domainName	VARCHAR	255	
connectionSpeed	VARCHAR	255	
isp	VARCHAR	255	
autonomousSystemNumber	INT	11	
asnOwner	VARCHAR	255	
proxyType	VARCHAR	255	
dma	VARCHAR	255	
isAnonymous	VARCHAR	255	
isAnonymousVpn	VARCHAR	255	
isHostingProvider	VARCHAR	255	
isPublicProxy	VARCHAR	255	
isTorExitNode	VARCHAR	255	
Status	VARCHAR	255	

Appendix B

Haversine Python Script

```
#!/usr/bin/env python
```

Jon Glas, jrg644ss@yahoo.com

```
import math
import sys
import csv
import itertools
import statistics

def main(arguments):
 #load data
 currentBin = ""
 ipList = []
 with open("iplist.txt", 'r') as infile:
  reader = csv.reader(infile)
  for row in reader:
   #check if bid is what we are already tracking. if it is append it to current list of ips
   if(currentBin == row[0]):
    ipList.append(row)
   elif(currentBin == ""): #check if it's the first iteration, if so we just start tracking the bin
    currentBin = row[0]
    ipList.append(row)
   else: # lastly it must be a brand new bin so we process the ipList and start tracking over
    print("calc now: " + row[0])
    calcDistance(ipList)
    ipList = []
    ipList.append(row)
    currentBin = row[0]
 # final bin needs to be calc'd still
 print("Final calc: " + row[0])
 calcDistance(ipList)
 infile.close()

def calcDistance(unorg_lst):
 dists = []
 # store the bin
 dists.append(unorg_lst[0][0])
```

```
#create a new list of just the lat/long combinations (strip out the ips)
for j in unorg_lst:
  del j[1]
  del j[0]
unorg_lst.sort()
# we remove ips that have dup lat/longs
lst = list(unorg_lst for unorg_lst,_ in itertools.groupby(unorg_lst))
#for i in range(0,len(lst)-2):
# start iterating the list and pop off one set of coords to compare to the rest of the list, repeat until list is empty
for i in range(0,1):
  for y in range(0,len(lst)-1):
    cur = lst.pop()
    # compare popped coords to the rest of the coords in the list
    for x in range(0,len(lst)):
      d = distance((float(cur[0]), float(cur[1])),(float(lst[x][0]), float(lst[x][1])))
      #print(d)
      dists.append(d)
#we then compute mean radius of all distances
row = [dists[0]]
del dists[0]
row.append(statistics.mean(dists))
write(row)


def write(row):
  with open('stats.csv', 'a') as writeFile:
    writer = csv.writer(writeFile)
    writer.writerow(row)
  writeFile.close()



# Haversine formula example in Python
# Author: Wayne Dyck
# example: print(distance((40.6469, -73.9344),(40.8881, -73.8414)))
def distance(origin, destination):
  lat1, lon1 = origin
```

```
    lat2, lon2 = destination
    radius = 3959 # miles

    dlat = math.radians(lat2-lat1)
    dlon = math.radians(lon2-lon1)
    a = math.sin(dlat/2) * math.sin(dlat/2) + math.cos(math.radians(lat1)) \
      * math.cos(math.radians(lat2)) * math.sin(dlon/2) * math.sin(dlon/2)
    c = 2 * math.atan2(math.sqrt(a), math.sqrt(1-a))
    d = radius * c
    return d

if __name__ == '__main__':
sys.exit(main(sys.argv[1:]))
```

Appendix C

Pearson Coefficient Formula

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{(n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2)(n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2)}}$$

Jon Glas, jrg644ss@yahoo.com