# Estimating Average Levels of Crime Reported within the Vicinity of a Metro Exit

Shawn Martin Tara Murphy Evan Rosenman Deeza-Mae Smith
Sean Wilson

July 22, 2015

**Abstract**

We utilize data from the Metropolitan Police Department (MPD) online database to create a finite population of property crimes occurring in the metropolitan D.C. area from January 1, 2014 through July 1, 2015. We employ clustering algorithms in the statistical package R to create 39 unique centroids defined by an individual or a cluster of Metro station exits and assign crimes to these locations. We then stratify this finite population by D.C.'s eight wards and employ Exact Optimal Allocation to construct a sample from each ward. Finally, we use statistical techniques to infer our primary parameter of interest: the average number of property crimes occurring within a 150 meter radius of a given metro exit on a weekend. We ultimately arrive at a 95% confidence interval of 0.2234 to 0.3247 property crimes per station exit per weekend.

# 1 Introduction

The choice of a place of work, residence, or leisure involves consideration of several factors, including convenience of transportation, proximity to restaurants or schools, and safety. This decision often involves trade-offs. For example, proximity to a Metro stop and a lively downtown area may come at the expense of a quiet neighborhood or low levels of crime. Many people who work or study within the District of Columbia (D.C.) consider these trade-offs when choosing where to reside within D.C.'s eight wards; however, potential residents may not know details of all the above mentioned qualities for different wards. In addition, it may be difficult to obtain reliable data and estimate statistics, including averages and totals, for all of these ward qualities. Under the assumption that the various characteristics of a ward are correlated with each other, estimation of one may aid in the inference of another.

The availability and convenience of public transportation is a particularly salient factor influencing the choice of residence in D.C. because of the high proportion of residents who use public transportation. The 2013 American Community Survey [1] estimates this figure to be 38.4% of D.C. residents who use public transportation (excluding taxis) as their primary means of transportation to work. Presumably, those who choose to use public transportation as their primary means of transportation will access the Metro or a bus at the closest station or stop. If information is known about the surrounding areas associated with the Metro stations or bus stops, then residents may use this information to augment their understanding of the area's related other qualities.

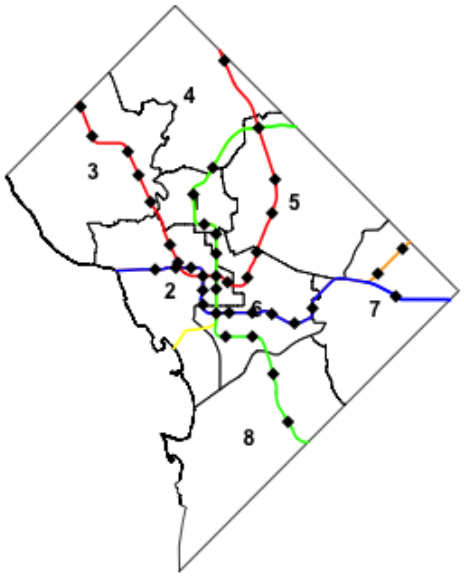
The availability of large public databases, including the MPD crime database and D.C. Open Data, allows data on the location of crimes committed and the location of Metro stations to be linked. In the sections that follow, we attempt to explore this relationship by calculating the average and total amount of crime that occurs in the vicinity of a Metro station. We hope the results presented in this report will aid the future study and estimation of other related ward-characteristics.

# 2 Population Definition & Primary Parameters

We focus on measuring the average level of crime associated with surrounding areas of Metro stations within D.C. There are 40 Metro stations within the metropolitan area of D.C. We define metropolitan D.C. as the area spanned by the eight wards that divide the city and are under the District of Columbia jurisdiction. The clear delineation of D.C. by ward is convenient for stratified sampling. In addition, because all Metro stations within the same

ward are confined to a specific geographic region, the characteristics of the area surrounding these stations is plausibly related. Wards typically vary across socioeconomic factors that can be correlated with crime, including income, education level, and housing prices. That is, we expect areas within wards to be similar across these auxiliary characteristics, and thus we might expect Metro stations in the same ward to experience similar levels of crime.

Figure 1: D.C. Metro Lines

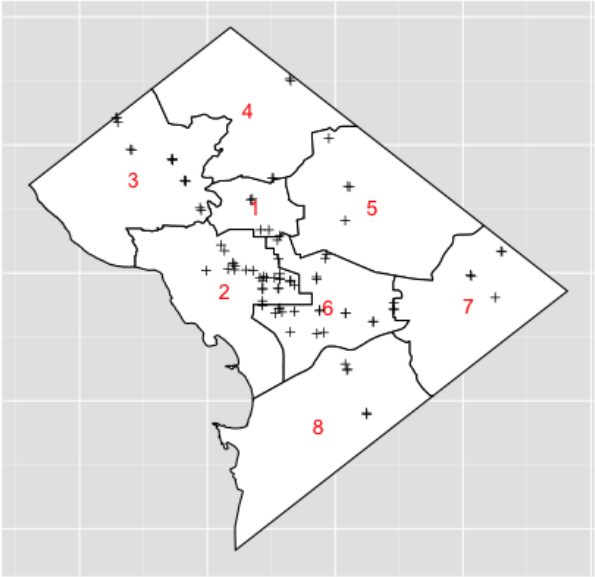



Each of the forty Metro stations in D.C. lies along one or more of the six Metro lines, shown above in Figure 1. Thirty-three of the forty Metro stations (roughly 83%) have more than one exit. For two such stations, these exits lie in different wards: the Shaw-Howard University station has exits in Wards 1 and 6, and the Smithsonian station has exits in Wards 2 and 6. Because these stations cannot be said to lie in a single ward, we redefine our locations of interest to be the 88 unique Metro station exits within D.C., shown in Figure 2.

We then associate crime records with station exits, which is achieved by drawing a circle of radius $r$ around each station exit and counting the number of crimes falling within that circle. We assume that a crime located within one block of a Metro station exit could reasonably be associated with that exit. Although the length of a block varies considerably, various sources [2] [3] [4] suggest that D.C. has, on average, ten to eleven blocks per mile (1.6 kilometers), which implies that an average block is approximately 146 - 160 meters long. We thus choose $r = 150$ meters.

Choosing this value of $r$ means that many of the circles would overlap, both in the case

Figure 2: Station Exits




of multiple exits for the same station and in the case of closely situated exits for different stations, e.g., Farragut North and Farragut West, which, at 171 meters apart, are the closest stations in D.C. Any crime falling in such an overlap would have to be either associated with each nearby station exit, or assigned to only one of them on the basis of some rule. To simplify this process, we cluster station exits by the following algorithm:

1. For every pair of station exits $\{(i, j) : i \neq j\}$, calculate the distance between $i$ and $j$.

2. If $i$ is closest to $j$, $j$ is closest to $i$, and the distance between $i$ and $j$ is less than $2r = 300$ meters, calculate the centroid of $i$ and $j$.

3. Let $k$ be the number of centroids calculated in step 2 plus the number of station exits not clustered in step 2. With $\{(i, j) : i + j = k, i \neq j\}$, repeat steps 1 and 2 until $k$ does not decrease from the previous iteration.

This algorithm produces 39 clusters of station exits, shown in Figure 3, for $r = 150$ meters. We are now able to assign each crime to at most one cluster.

The MPD [5] categorizes offenses as 'violent' or 'property' crimes, shown in Table 1. Although both categories of crime are of general interest, we observe a pattern wherein all homicides were reported precisely at midnight, which appears to be an artifact of the

reporting process. To avoid any bias this might introduce, we restrict our investigation to property crimes.

Figure 3: Metro Stations Exit Clusters

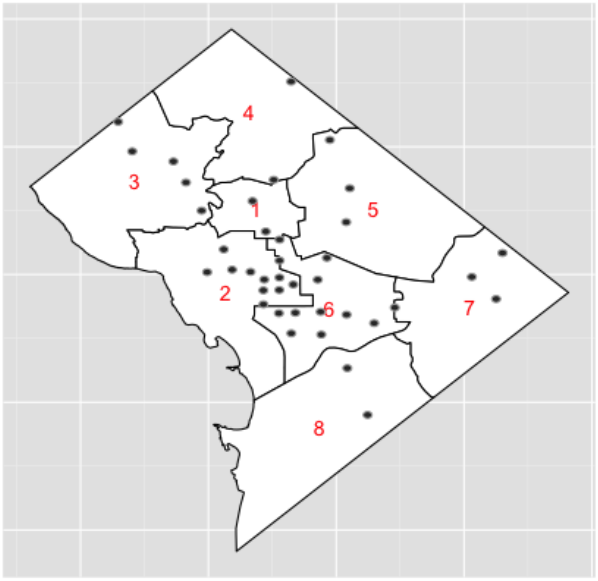



Table 1: Offense by Category

| **Category** | **Offense** |
|---|---|
| Violent | Assault w/ dangerous weapon |
| Violent | Homicide |
| Violent | Robbery |
| Violent | Sex Abuse |
| Property | Arson |
| Property | Burglary |
| Property | Motor vehicle theft |
| Property | Theft from auto |
| Property | Theft/other |

Weekend Metro ridership patterns are different from weekday patterns [6]. We assume that weekend ridership includes a higher proportion of discretionary travel (i.e., non-commuter travel) and that focusing our attention on weekend crime patterns might, therefore, be of greater interest for a rider taking such a 'marginal' trip. Thus, we define our unit of

analysis to be the combination of Metro station exit cluster and weekend. We opt to look at all weekends in the last twelve months (from July 1, 2014 to July 1, 2015), so that our population consists of the 2,028 combinations of the 52 weekends and 39 station exit clusters.

Given this population, our primary parameter of interest, $\bar{y}_U$, is the mean number of property crimes reported on a weekend within a 150-meter radius of a clustered Metro station exit. A secondary parameter of interest is the total number of such crimes, $t$.

# 3 Data Collection Instrument & Sampling Frame

We extract station exit location data from the D.C. Open Data website [7]. This data set consists of the locations of the 88 station exits, given as latitude and longitude using the WGS84 datum (see Appendix, Section 8), and auxiliary information about the location, such as station name and street address. To these stations, we apply the clustering algorithm described in Section 2, resulting in 39 clusters. We also extract ward location data from the D.C. Open Data website [8], which we use to assign each cluster to a ward. Then around the centroids of each cluster, we draw a circle with radius of size 150 meters. We use the R packages `dplyr`, `geosphere`, `rgdal`, and `rgeos` to read and manipulate the data.

We then extract from the MPD crime database all crimes reported within D.C. between January 1, 2014 and July 1, 2015. This data set consists of 49,548 records and includes the date and time each crime was reported, the location of each crime, the type of offense, and other auxiliary information. Using the R package `lubridate`, we filter the data set to include only property crimes reported during a weekend (i.e., those reported on a Saturday or a Sunday.)

The MPD reports the location of a crime as "approximated to the center of the street block", [5]. Our choice of $r = 150$ meters thus ensures that a Metro station located on the corner of a block (i.e., as far as possible from the center of the block) would still be associated with any crimes reported on that block. We transform crime location, which is given in terms of the Maryland State Plane meters NAD83 datum, to the WGS84 datum. We then use the R package `sp` to determine which crimes fall within the 150-meter circles plotted around the 39 clustered station exit centroids.

For each station exit cluster and weekend pair, we count the number of property crimes that had been reported within 150 meters of that cluster on that weekend. The resulting data set becomes our sampling frame, and consists of identifiers for the station exit cluster and weekend, the ward number, and the number of reported crimes. Table 2 shows a few

observations from the sampling frame. As described in Section 2, the sampling frame consists of 2,028 observations of the 39 station exit clusters on each of the 52 weekends in the 12-month period July 1, 2014 - July 1, 2015.

Table 2: Sampling Frame

| WKND_ID | WKND_START | WKND_END | CTRD_ID | WARD | CTRD_LAT | CTRD_LONG | CATEGORY | NUM_CRIMES |
|---|---|---|---|---|---|---|---|---|
| 37 | 9/13/14 | 9/14/14 | 1 | 7 | 38.9527 | -77.002 | property | 1 |
| 38 | 9/20/14 | 9/21/14 | 1 | 7 | 38.9527 | -77.002 | property | 1 |
| 43 | 10/25/14 | 10/26/14 | 1 | 7 | 38.9527 | -77.002 | property | 1 |
| 45 | 11/8/14 | 11/9/14 | 1 | 7 | 38.9527 | -77.002 | property | 1 |
| 47 | 11/22/14 | 11/23/14 | 1 | 7 | 38.9527 | -77.002 | property | 1 |

# 4 Sampling & Estimation Plan

## 4.1 Pilot Survey & Determining $n$

Given our constructed dataset detailing the number of property crimes within 150 meters of a centroid of metro station exits on weekends, we proceed to stratify our sample by ward. We seek to determine the lowest sample size ($n$) and stratum sample sizes ($n_h$ for $h = 1$, 2,...,8) to achieve a desired accuracy of our estimate. We know $\bar{y}_{str}$ is an unbiased estimator of $\bar{y}_U$ (i.e., $\mathrm{E}[\bar{y}_{str}] = \bar{y}_U$) and we use $\bar{y}_{str}$ to estimate $\bar{y}_U$, the true average number of crimes in our population.

We seek a sample size $n$ that will ensure our estimate, $\bar{y}_{str}$, lies within 0.05 crimes of $\bar{y}_U$ with probability 0.95. In other words, we want a sample size such that $P(\mid \bar{y}_{str} - \bar{y}_U \mid < 0.05) = 0.95$. Given this constraint, we proceed in the following steps:

1. Estimate the population standard deviation, $S_h$, of each stratum (ward) using data from the pilot survey, for $h = 1$, 2...8;

2. Conduct Neyman Allocation to calculate an initial estimate of the total number of units, $n$, to sample across all eight wards;

3. Employ Exact Optimal Allocation with this estimated $n$. Then, in order to check that our sample size is adequate for the desired accuracy, estimate the variance of our estimator $V(\bar{y}_{str})$ by calculating $\hat{V}(\bar{y}_{str})$ and check if $\hat{V}(\bar{y}_{str}) < \frac{e^2}{z^2_{\frac{\alpha}{2}}} = \frac{(0.05)^2}{1.96^2} = 0.00065$:

    - If so, move forward with these allocations.

- If not, repeatedly increment the total number to sample $n$ and re-check this condition until it is met.

We conduct a pilot survey through which we calculate the sample variance, $s_h^2$, and sample standard deviation, $s_h$, for each ward. While we know that $s_h$ is a biased estimator of $S_h$, we aim to minimize this bias by taking a large number of units to be part of our pilot survey. We treat $s_h$ as our $S_h$, the population standard deviation, for the purposes of sample allocation.

We select a subset of our data consisting of all crimes from January 1, 2014 to June 30, 2014 to comprise our pilot survey. This time period consists of 26 weekends $\times$ 39 metro exit clusters for a total of 1,014 observations. We use the entirety of these observations to calculate our estimates, so $n$ = 1,014. We calculate $s_h$ within each stratum leveraging the standard formula:

$$s_h = \sqrt{\frac{\sum_{i=1}^{n_h}(y_i - \bar{y}_h)^2}{n_h - 1}} \tag{1}$$

We arrive at the results in Table 3.

Table 3: Pilot Survey $s_h$ Values, by Ward

| **ward** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
|---|---|---|---|---|---|---|---|---|
| $s_h$ | 1.19245 | 0.61929 | 0.53215 | 0.55606 | 0.43891 | 0.42396 | 0.54852 | 0.23544 |

One important caveat to this pilot survey design is that we are only looking at a subset of of the calendar year. If there are strong seasonal patterns in crime frequency, our pilot survey may not be representative of yearly trends in crime level. This may introduce selection bias into our estimates of $S_h$; however, with half the year represented, we believe this selection bias will be relatively small.

To allocate $n$ to $n_h$, we first utilize Neyman Allocation with costs assumed $c_1 = c_2 = ... = c_H$ and use the sample variances, $s_h^2$, obtained from the pilot survey as unbiased estimators for $S_h^2$ :

$$n = \frac{\sum_{h=1}^{H} N_h^2 \frac{S_h^2}{w_h^*}}{(\sum_{h=1}^{H} N_h S_h^2 + N^2 \frac{e^2}{z_{\frac{\alpha}{2}}^2})}, \quad where \quad w_h^* = \frac{N_h S_h}{\sum_{i=1}^{H} N_i S_i} \tag{2}$$

Applying the ceiling function to the results from Neyman allocation, we determine an overall sample size of $n = 364$ will allow us to estimate $\bar{y}_U$ with $\bar{y}_{str}$ such that $P(|\ \bar{y}_{str}$ - $\bar{y}_U\ | < e) = 0.95$, where $e = 0.05$.

## 4.2 Exact Optimal Allocation

With an initial sample size of $n = 364$, we use Exact Optimal Allocation Algorithm II to calculate the sample size for each ward. We initially assign two units to each stratum, then compute the priority matrix of 8 x 348 values. The first few columns of this matrix is depicted below in Figure 4.

Figure 4: Exact Optimal Allocation Matrix

| | [,1] | [,2] | [,3] | [,4] | [,5] | [,6] | [,7] | [,8] | [,9] | [,10] |
|---|---|---|---|---|---|---|---|---|---|---|
| [1,] | 50.628985 | 35.800099 | 27.730637 | 22.641970 | 19.135958 | 16.572225 | 14.615329 | 13.072348 | 11.824383 | 10.794136 |
| [2,] | 144.616015 | 102.258965 | 79.209454 | 64.674248 | 54.659716 | 47.336703 | 41.747048 | 37.339695 | 33.775025 | 30.832238 |
| [3,] | 56.484889 | 39.940848 | 30.938048 | 25.260810 | 21.349281 | 18.489020 | 16.305783 | 14.584336 | 13.192028 | 12.042619 |
| [4,] | 23.609016 | 16.694095 | 12.931190 | 10.558273 | 8.923369 | 7.727864 | 6.815336 | 6.095822 | 5.513878 | 5.033459 |
| [5,] | 27.952650 | 19.765508 | 15.310297 | 12.500805 | 10.565109 | 9.149653 | 8.069235 | 7.217343 | 6.528333 | 5.959525 |
| [6,] | 90.002717 | 63.641531 | 49.296518 | 40.250439 | 34.017829 | 29.460305 | 25.981546 | 23.238602 | 21.020106 | 19.188644 |
| [7,] | 46.578187 | 32.935752 | 25.511924 | 20.830398 | 17.604900 | 15.246291 | 13.445964 | 12.026436 | 10.878321 | 9.930503 |
| [8,] | 9.996078 | 7.068294 | 5.475077 | 4.470382 | 3.778162 | 3.271984 | 2.885619 | 2.580976 | 2.334581 | 2.131171 |

We choose the 348 highest values from the matrix in order to determine the prospective number of units to sample for each ward. Then, we use Equation (3) to estimate $\hat{V}(\bar{y}_{str})$ under these allocations. We seek $\hat{V}(\bar{y}_{str}) < \frac{e^2}{z^2_{\frac{\alpha}{2}}} = \frac{(0.05)^2}{1.96^2} = 0.00065$. This would indicate that this allocation of $n$ to $n_h$ should result in a sample average within 0.05 crimes of the true average, with a probability of 0.95.

$$\hat{V}(\bar{y}_{str}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \hat{V}(\bar{y}_h) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{N_h - n_h}{N_h} \frac{s_h^2}{n_h} \tag{3}$$

We find this condition to be satisfied by computing $\hat{V}(\bar{y}_{str}) = 0.000649 < 0.00065$. Hence, we proceed with this allocation of $n_h$. The allocation of the sample to the wards is displayed in Table 4.

Under stratified random sampling, the probability of selection of the $j^{th}$ unit in the $h^{th}$ ward is $\frac{n_h}{N_h}$ and displayed for each ward in Table 4. The probability of each possible stratified random sample is calculated using Expression (4) below. The probability is extremely small and is approximately equal to $P(\text{each stratified random sample}) = 1.33 \times 10^{-396}$.

Table 4: Allocation of Sample to Ward, $n_h$

| $h$ | **Sample size** ($n_h$) | **Population size** ($N_h$) | **Selection Probability** |
|---|---|---|---|
| 1 | 41 | 104 | 0.39 |
| 2 | 116 | 572 | 0.2 |
| 3 | 46 | 260 | 0.18 |
| 4 | 19 | 104 | 0.18 |
| 5 | 23 | 156 | 0.15 |
| 6 | 73 | 520 | 0.14 |
| 7 | 38 | 208 | 0.18 |
| 8 | 8 | 104 | 0.08 |

$$\frac{1}{\binom{N_1}{n_1}\binom{N_2}{n_2}\binom{N_3}{n_3}\binom{N_4}{n_4}\binom{N_5}{n_5}\binom{N_6}{n_6}\binom{N_7}{n_7}\binom{N_8}{n_8}} \tag{4}$$

# 5 Sampling Results

## 5.1 Estimating $\bar{y}_U$

Now that the sample size is determined, we can sample from our dataset by using R's `sample` function, which allows us to select integers pseudo-randomly from a given range. For our purposes, this function can be used to select units from within each ward to include in our sample. We can thus estimate the average number of crimes that occurred both within each ward and throughout D.C. on weekends from July 1, 2014 to July 1, 2015 within 150 meters of the station exit centroids. We find $\bar{y}_h$ for wards $h = 1, 2, ..., 8$ and $\bar{y}_{str}$ for the stratified sample of D.C. to estimate $\bar{y}_{hU}$ and $\bar{y}_U$, respectively. This is calculated using Equations (5) and (6) and results are displayed in Table 5. Computing this weighted mean, we estimate $\bar{y}_U$ with $\bar{y}_{str} = 0.2740$ . Figure 5 shows $\bar{y}_h$ by ward.

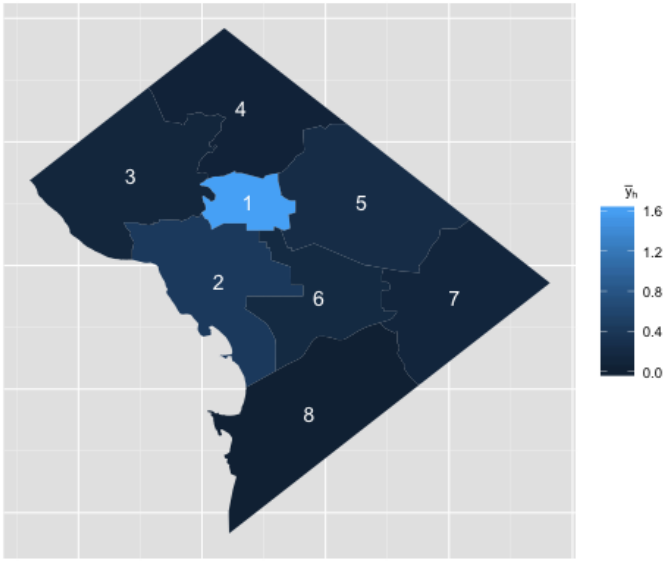
$$\bar{y}_h = \sum_{j=1}^{n_h} \frac{y_{hj}}{n_h} \tag{5}$$

$$\bar{y}_{str} = \frac{N_1}{N}\bar{y}_1 + \frac{N_2}{N}\bar{y}_2 + ... + \frac{N_h}{N}\bar{y}_h + ... + \frac{N_H}{N}\bar{y}_H \tag{6}$$

Table 5: $\bar{y}_h$ by Ward

| $h$ | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
|---|---|---|---|---|---|---|---|---|
| $\bar{y}_h$ | 1.71 | 0.39 | 0.39 | 0.32 | 0.17 | 0.27 | 0.11 | 0.00 |

Figure 5: Sample Mean by Ward




The variability of our stratified sample is calculated using Equation (7) and Equation (8) with results shown in Table 6.

$$\hat{V}(\bar{y}_h) = \frac{N_h - n_h}{N_h}\frac{s_h^2}{n_h} \tag{7}$$

$$\hat{V}(\bar{y}_{str}) = \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2 \hat{V}(\bar{y}_h) = \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2 \frac{N_h - n_h}{N_h}\frac{s_h^2}{n_h} \tag{8}$$

Table 6: Sample Variation by Ward

| $h$ | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
|---|---|---|---|---|---|---|---|---|
| $s_h^2$ | 3.0600 | 0.5700 | 0.6400 | 0.4500 | 0.1500 | 0.2900 | 0.1000 | 0.0000 |
| $\hat{V}(\bar{y}_h)$ | 0.0452 | 0.0039 | 0.0115 | 0.0194 | 0.0056 | 0.0034 | 0.0021 | 0.0000 |

We can thus estimate $V(\bar{y}_{str})$ with $\hat{V}(\bar{y}_{str}) = 0.0006674$. Using Equation (9), we find a 95% confidence interval for $\bar{y}_U$ to be (0.2234, 0.3247).

$$\left(\bar{y}_{str} - z_{\frac{\alpha}{2}}\sqrt{\hat{V}(\bar{y}_{str})},\ \ \bar{y}_{str} + z_{\frac{\alpha}{2}}\sqrt{\hat{V}(\bar{y}_{str})}\right) \tag{9}$$

## 5.2 Estimating t

We also estimate the total number of crimes that occurred both within each ward and throughout D.C. on weekends from July 1, 2014 to July 1, 2015 within 150 meters of the station exit centroids, $t_h$ and $t$, with estimators $\hat{t}_h$ and $\hat{t}_{str}$ for $h = 1, 2,...,8$ using Equations (10) and (11), respectively. First, we consider sample sizes. We note that, assuming the same total number of samples, $n = 364$, Exact Optimal Allocation will choose identical allocations by stratum for sample totals as for sample means. Moreover, we recall that $\mathrm{V}(\hat{t}_{str}) = N^2 V(\bar{y}_{str})$, which implies that $z_{\frac{\alpha}{2}}\sqrt{V(\hat{t}_{str})} = N z_{\frac{\alpha}{2}}\sqrt{V(\bar{y}_{str})} = Ne$. Since our original $e = 0.05$, this implies that our estimate, $\hat{t}_{str}$, will be within $0.05 \cdot 2028 = 101.4$ of the true value of $t$ with probability 95%. We consider this acceptable, and hence proceed with the exact same sample as when calculating $\bar{y}_{str}$. Results are shown in Table 7.

$$\hat{t}_h = N_h \bar{y}_h \tag{10}$$

$$\hat{t}_{str} = N\bar{y}_{str} = \hat{t}_1 + ... + \hat{t}_h + ... + \hat{t}_H = N_1\bar{y}_1 + ... + N_h\bar{y}_h + ... + N_H\bar{y}_H \tag{11}$$

Table 7: Sample Totals, $\hat{t}_h$, by Ward

| $h$ | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
|---|---|---|---|---|---|---|---|---|
| $\hat{t}_h$ | 152.20 | 172.59 | 33.91 | 54.74 | 13.57 | 106.85 | 21.89 | 0.00 |

Our estimate of $t$ is $\hat{t}_{str} = 555.74$, with estimated variance $\hat{V}(\hat{t}_{str}) = 2744.74$, as calculated using Equation (12). Use of Equation (13) leads to an overall 95% confidence interval for $t$ of (453.06, 658.42) total crimes.

$$\hat{V}(\hat{t}_{str}) = \sum_{h=1}^{H} N_h^2 \hat{V}(\bar{y}_h) = \sum_{h=1}^{H} N_h^2 (\frac{N_h - n_h}{N_h})\frac{s_h^2}{n_h} \tag{12}$$

$$\left(\hat{t} - z_{\frac{\alpha}{2}}\sqrt{\hat{V}(\hat{t})},\ \ \hat{t} + z_{\frac{\alpha}{2}}\sqrt{\hat{V}(\hat{t})}\right) \tag{13}$$

# 6 Comparison: Sample vs. Population Statistics

## 6.1 Sample Statistic Accuracy

Because we have access to the underlying data set of property crime numbers for all 2,028 weekend-centroid pairs, we are able to derive a number of further insights from the data. Since the samples are identical for our estimation of $\bar{y}_{str}$ and $\hat{t}_{str}$, and these estimates differ only by a constant, we focus our attention on $\bar{y}_{str}$. We are able to directly compute the true value of $\bar{y}_U$ and verify that our confidence interval covers this value. It turns out that the true value is 0.286, so our confidence interval of (0.2234, 0.3247) indeed covers this value. We are also able to compare the values of $\bar{y}_h$ against the true population parameters $\bar{y}_{hU}$. These values are given in Tables 8 and 9 below and, as we can see, they align closely.

Table 8: $\bar{y}_h$ by Ward

| $h$ | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
|---|---|---|---|---|---|---|---|---|
| $\bar{y}_h$ | 1.71 | 0.39 | 0.39 | 0.32 | 0.17 | 0.27 | 0.11 | 0.00 |

Table 9: $\bar{y}_{hU}$ by Ward

| $h$ | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
|---|---|---|---|---|---|---|---|---|
| $\bar{y}_{hU}$ | 1.44 | 0.30 | 0.30 | 0.38 | 0.13 | 0.18 | 0.12 | 0.01 |

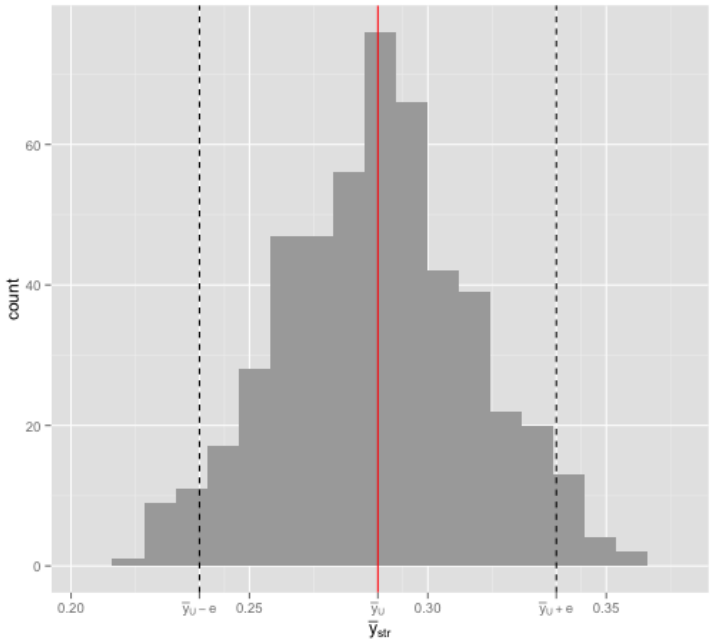
We are able to explore the overall accuracy of our estimates using stratified samples. To do so, we repeatedly sample from the data, changing the random seed each time such that a new sample would be collected within each ward. We do this 500 times and look at the distribution of our 500 different values of $\bar{y}_{str}$. A histogram of these estimates is displayed in Figure 6 below. The red line represents the population mean of 0.286, while the dashed black lines are values above and below the population mean by $e = 0.05$.

As we can see, the distribution:

- appears Gaussian, as the Central Limit Theorem implies;
- is centered almost exactly at the true population mean (the mean of the 500 estimates is 0.2856);

- the vast majority of the estimates fall within $e = 0.05$ of the true population mean, as desired.

Figure 6: Sampling Distribution of $\bar{y}_{str}$




It turns out that 92.2% of our estimates fall within $e = 0.05$, just shy of the 95% target. This may be due to the fact that variance is estimated using only the first six months of 2014, and there may be some seasonal patterns in crime.

## 6.2 Alternative Sampling Schemes

Lastly, we seek to explore two other sampling schemes:

- simple random sampling
- stratified random sampling, where each stratum is defined as a single centroid, rather than a ward

### 6.2.1 Sample Size Comparison

We are interested in learning how these sampling schemes compare against our chosen method of stratifying based on ward. First, we compute the required sample size under

each alternative sampling scheme such that our estimate will be within $e = 0.05$ of the true population mean with 95% probability.

For Simple Random Sampling, we perform our computations using Equation (14) :

$$n = \frac{N\sigma^2}{\left((N-1)\frac{e^2}{z^2_{\frac{\alpha}{2}}} + \sigma^2\right)} \tag{14}$$

This yields a required sample size of 441.4 units, which we round up to 442. This is more than 20% larger than the number of required sample units under stratified random sampling by ward, 364.

For stratified random sampling by centroid, we use the same procedure as in Section 4.2 in order to compute our sample allocations under Exact Optimal Allocation. In this case, when we use our pilot survey data to get to an estimate of the variance by station, we wind up with a number of stations for which there were no property crimes reported. Our variance estimate is hence 0.00; however, because we are using Exact Optimal Allocation Algorithm II, we still allocate two sampling units to each of these stations.

Our algorithm increments the number of sampling units until $\hat{V}(\bar{y}_{str}) < \frac{e^2}{z^2_{\frac{\alpha}{2}}} = \frac{(0.05)^2}{1.96^2} = 0.00065$. This resulting allocations are shown below in Table 10.

Table 10: Allocation of Sample to Centroid, $n_h$

| **Stratum (Centroid)** | **Sample size $n_h$** | **Stratum (Centroid)** | **Sample size $n_h$** |
|---|---|---|---|
| 1 | 3 | 21 | 4 |
| 2 | 9 | 22 | 3 |
| 3 | 10 | 23 | 9 |
| 4 | 7 | 24 | 6 |
| 5 | 3 | 25 | 4 |
| 6 | 11 | 26 | 17 |
| 7 | 2 | 27 | 4 |
| 8 | 12 | 28 | 9 |
| 9 | 3 | 29 | 10 |
| 10 | 5 | 30 | 2 |
| 11 | 6 | 31 | 2 |
| 12 | 9 | 32 | 2 |
| 13 | 4 | 33 | 2 |
| 14 | 7 | 34 | 3 |
| 15 | 8 | 35 | 3 |
| 16 | 9 | 36 | 2 |
| 17 | 5 | 37 | 9 |
| 18 | 3 | 38 | 11 |
| 19 | 5 | 39 | 3 |
| 20 | 2 | | |

The sum of all units across all centroids is just 228 – a nearly 40% reduction from the number required when stratifying by ward! Hence, if there were substantial costs associated with sampling each unit, this would be far more efficient than stratifying by ward.

Table 11 below summarizes these results.

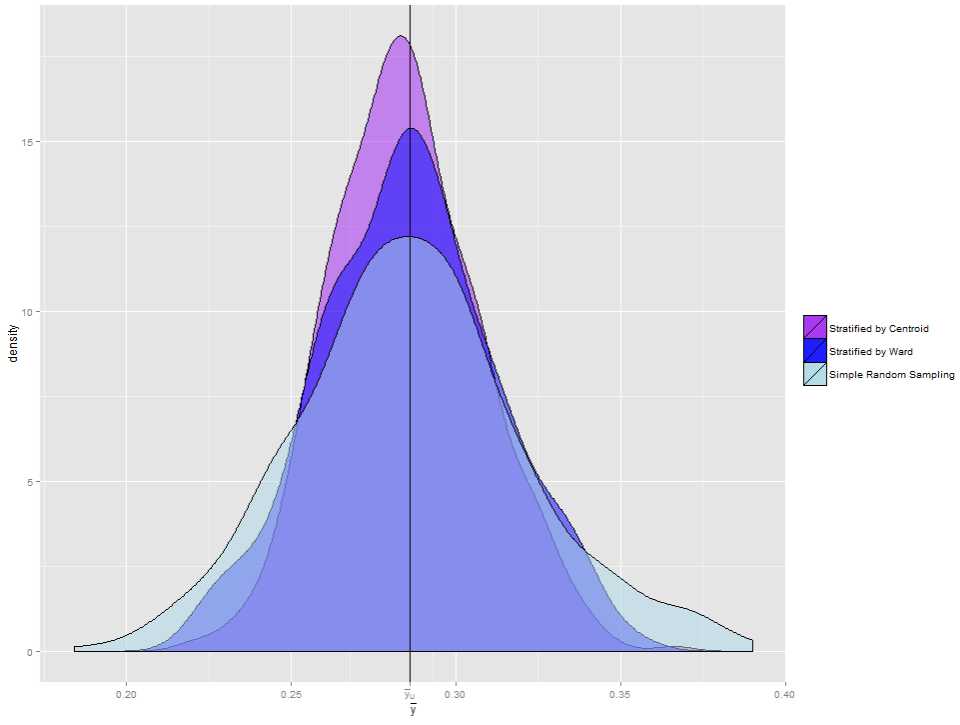
Table 11: Required Sample Sizes under 3 Sampling Schemes

| Sampling Scheme | $n$ |
|---|---|
| SRS | 442 |
| Stratified by Ward | 364 |
| Stratified by Centroid | 228 |

### 6.2.2 Distribution Comparison

As a second way of comparing these sampling schemes, we conduct 500 iterations of resampling for each method, using a total sample size of $n$=364, the same number of total samples required for stratified random sampling by ward. We then generate a density plot, depicted in Figure 7 below, of the estimates for each method:

Figure 7: Density Estimate of Sampling Distribution of $\bar{y}$



This plot shows exactly what we expect: the three distributions appear to be roughly normal and to share the same mean. The distribution of estimates from stratified random sampling by centroid is narrowest and most peaked near the true population mean. The stratified random sampling by ward distribution is displayed in the middle, and the simple random sampling distribution is the widest. This indicates that the variance of our estimator is meaningfully reduced when stratifying by ward and dramatically reduced when stratifying by centroid. By stratifying with increasing granularity, we are able to get more accurate estimates of the true population mean – and narrower confidence intervals – than we would be able to achieve using the same number of sample units under simple random sampling.

# 7 Conclusion

Our estimate of $\bar{y}_{str} = 0.2740$ implies that slightly more than one crime is reported within 150 meters of a (clustered) Metro station exit in D.C. over a four-week period. Ward 1 has the highest sample mean (1.71 crimes), while Ward 8 has the lowest sample mean (0 crimes). Sample means for the other wards are similar in magnitude to $\bar{y}_{str}$. Because we have the population, we are able to compare our estimate to the population parameter $\bar{y}_U$, noting as in Section 6.1 that our estimate differs from the true value by 0.012 crimes. We also confirm empirically that thoughtful stratification and sample allocation allow us to attain a lower variance for our estimator than under simple random sampling for a fixed sample size.

Our estimate of the average number of crimes outside a Metro station on a weekend may be affected by several sources of bias:

1. We opt to classify a crime as 'weekend' or 'non-weekend' based on when the crime is reported, which will tend to lag when the crime is actually committed. This lag likely means that the set of crimes classified as 'weekend' is different in both number and distribution of attributes from the set of crimes actually committed on a weekend.

2. Because not all crimes are reported, we may tend to underestimate the number of crimes that are actually committed.

3. Our source for crime data is the Metropolitan Police Department, but the MPD is not the only law enforcement agency operating in D.C. Any crimes reported to other law enforcement agencies, such as the US Park Police or US Secret Service, might not be included in our data set.

4. Reported crime locations are approximated to the center of the block, but not all blocks in D.C. have the same length. Although the circles drawn around the clustered station exits have a radius roughly equal to the length of an entire block, there may be cases in which the length of a block close to a station is such that this radius will fail to capture the center of the block. In such a case, we would not associate any crimes reported for that block with the station.

5. We select our sample using the R function `sample`, which relies on R's random number generator. This process is not truly random, but pseudo-random. Although this is likely sufficient for our purposes, it is nevertheless a possible source of bias.

Our investigation attempts to estimate the average level of criminal activity around a Metro station exit on the weekend. Further research should address the sources of potential bias discussed above, and should consider additional variables that may plausibly affect criminal activity, including station ridership, proximity to police stations or other locations of interest, availability of parking, and connections to other modes of transportation.

# 8 Appendix

Table 12: Station Exits with Cluster and Location

| Ward | Cluster ID | Station Exit Name | Longitude | Latitude |
|---|---|---|---|---|
| 1 | 16 | U St \| African Amer Civil \| War Mem \| Cardozo | -77.025964 | 38.916794 |
| 1 | 16 | U St \| African Amer Civil \| War Mem \| Cardozo | -77.029124 | 38.916831 |
| 1 | 26 | Columbia Heights | -77.032430 | 38.928876 |
| 1 | 26 | Columbia Heights | -77.033032 | 38.928619 |
| 2 | 3 | Mt Vernon Sq \| 7th St \| Convention Cen | -77.022150 | 38.905436 |
| 2 | 4 | Foggy Bottom \| GWU | -77.050465 | 38.900903 |
| 2 | 8 | Dupont Circle | -77.043319 | 38.908683 |
| 2 | 8 | Dupont Circle | -77.044618 | 38.910946 |
| 2 | 17 | McPherson Sq | -77.032134 | 38.900883 |
| 2 | 17 | McPherson Sq | -77.034857 | 38.901162 |
| 2 | 20 | Federal Triangle | -77.028658 | 38.893810 |
| 2 | 20 | Federal Triangle Elevator | -77.028316 | 38.893777 |
| 2 | 21 | Archives \| Navy Mem | -77.022440 | 38.893944 |
| 2 | 21 | Archives \| Navy Mem Elevator | -77.022081 | 38.893877 |
| 2 | 33 | Smithsonian | -77.028457 | 38.887255 |
| 2 | 33 | Smithsonian | -77.028437 | 38.889078 |
| 2 | 33 | Smithsonian Elevator | -77.028536 | 38.887757 |
| 2 | 35 | Judiciary Sq | -77.015987 | 38.895322 |
| 2 | 35 | Judiciary Sq | -77.017525 | 38.897060 |
| 2 | 35 | Judiciary Sq NE Elevator | -77.017453 | 38.896830 |
| 2 | 35 | Judiciary Sq SW Elevator | -77.017657 | 38.896654 |
| 2 | 37 | Gallery Pl \| Chinatown | -77.021567 | 38.897544 |
| 2 | 37 | Gallery Pl \| Chinatown | -77.023781 | 38.898166 |
| 2 | 37 | Gallery Pl \| Chinatown | -77.021656 | 38.899640 |
| 2 | 37 | Gallery Pl \| Chinatown Elevator | -77.021749 | 38.897845 |
| 2 | 38 | Metro Center | -77.026817 | 38.898164 |
| 2 | 38 | Metro Center | -77.027894 | 38.898527 |
| 2 | 38 | Metro Center | -77.029452 | 38.898141 |
| 2 | 38 | Metro Center | -77.028376 | 38.897128 |
| 2 | 39 | Farragut North | -77.039834 | 38.903930 |

Table 12: Station Exits with Cluster and Location

| Ward | Cluster ID | Station Exit Name | Longitude | Latitude |
|---|---|---|---|---|
| 2 | 39 | Farragut North | -77.040130 | 38.903590 |
| 2 | 39 | Farragut North | -77.039180 | 38.902781 |
| 2 | 39 | Farragut West | -77.041947 | 38.901481 |
| 2 | 39 | Farragut West | -77.039345 | 38.901252 |
| 3 | 14 | Tenleytown \| AU | -77.079462 | 38.948079 |
| 3 | 14 | Tenleytown \| AU | -77.080088 | 38.948284 |
| 3 | 15 | Woodley Park Zoo \| Adams Morgan | -77.052383 | 38.924510 |
| 3 | 15 | Woodley Park Zoo \| Adams Morgan Elevator | -77.052876 | 38.925457 |
| 3 | 29 | Friendship Heights | -77.085765 | 38.960453 |
| 3 | 29 | Friendship Heights | -77.085402 | 38.960748 |
| 3 | 29 | Friendship Heights Elevator | -77.085006 | 38.958902 |
| 3 | 30 | Van Ness \| UDC | -77.063526 | 38.944556 |
| 3 | 30 | Van Ness \| UDC | -77.063996 | 38.944476 |
| 3 | 30 | Van Ness \| UDC Elevator | -77.063581 | 38.943970 |
| 3 | 34 | Cleveland Park | -77.059007 | 38.936146 |
| 3 | 34 | Cleveland Park | -77.058537 | 38.935985 |
| 3 | 34 | Cleveland Park Elevator | -77.058437 | 38.935777 |
| 4 | 13 | Takoma | -77.017416 | 38.975104 |
| 4 | 13 | Takoma Elevator | -77.017824 | 38.975889 |
| 4 | 28 | Georgia Ave \| Petworth | -77.024715 | 38.937129 |
| 4 | 28 | Georgia Ave \| Petworth | -77.024161 | 38.936951 |
| 5 | 1 | Fort Totten | -77.002491 | 38.952654 |
| 5 | 2 | Rhode Island Ave | -76.996075 | 38.920478 |
| 5 | 9 | Brookland \| CUA | -76.995045 | 38.933834 |
| 5 | 9 | Brookland \| CUA | -76.994368 | 38.933630 |
| 6 | 5 | Waterfront | -77.017558 | 38.876962 |
| 6 | 7 | Federal Center SW | -77.015962 | 38.885002 |
| 6 | 11 | Shaw \| Howard Univ | -77.022598 | 38.912781 |
| 6 | 11 | Shaw \| Howard Univ | -77.021684 | 38.914506 |
| 6 | 12 | NoMa-Gallaudet U | -77.003430 | 38.907287 |
| 6 | 12 | NoMa-Gallaudet U | -77.003829 | 38.905718 |
| 6 | 19 | Union Station | -77.007202 | 38.897492 |

Table 12: Station Exits with Cluster and Location

| Ward | Cluster ID | Station Exit Name | Longitude | Latitude |
|---|---|---|---|---|
| 6 | 19 | Union Station | -77.007229 | 38.898400 |
| 6 | 22 | Capitol South | -77.006055 | 38.885570 |
| 6 | 22 | Capitol South Elevator | -77.006076 | 38.885181 |
| 6 | 23 | Eastern Market | -76.995815 | 38.884253 |
| 6 | 23 | Eastern Market Elevator | -76.996040 | 38.884261 |
| 6 | 24 | Potomac Ave | -76.985174 | 38.880847 |
| 6 | 24 | Potomac Ave Elevator | -76.985108 | 38.881071 |
| 6 | 25 | Navy Yard-Ballpark | -77.004497 | 38.876677 |
| 6 | 25 | Navy Yard-Ballpark | -77.007133 | 38.876278 |
| 6 | 36 | LEnfant Plaza | -77.021577 | 38.886255 |
| 6 | 36 | LEnfant Plaza | -77.023485 | 38.884253 |
| 6 | 36 | LEnfant Plaza | -77.020828 | 38.884689 |
| 6 | 36 | LEnfant Plaza Elevator | -77.022100 | 38.885985 |
| 7 | 6 | Benning Road | -76.937449 | 38.890405 |
| 7 | 10 | Deanwood | -76.934723 | 38.908268 |
| 7 | 10 | Deanwood | -76.935172 | 38.908561 |
| 7 | 18 | Minnesota Ave | -76.946751 | 38.898763 |
| 7 | 18 | Minnesota Ave | -76.947125 | 38.899311 |
| 7 | 31 | Stadium \| Armory | -76.977046 | 38.885663 |
| 7 | 31 | Stadium \| Armory | -76.977065 | 38.888291 |
| 7 | 31 | Stadium \| Armory Elevator | -76.977154 | 38.885929 |
| 8 | 27 | Congress Heights | -76.987620 | 38.844655 |
| 8 | 27 | Congress Heights | -76.987778 | 38.845398 |
| 8 | 32 | Anacostia | -76.995070 | 38.862298 |
| 8 | 32 | Anacostia | -76.996023 | 38.864436 |
| 8 | 32 | Anacostia | -76.995466 | 38.862169 |

# References


[1] U.S. Census Bureau. *"Commuting Characteristics by Sex: 2009-2013 American Community Survey 5-Year Estimates"*. 2013. URL: http://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk (visited on 07/21/2015).

[2] *"How far is a block in D.C."*. 2007. URL: http://www.tripadvisor.com/ShowTopic-g28970-i40-k1338560-How_far_is_a_block_in_DC-Washington_DC_District_of_Columbia.html (visited on 07/19/2015).

[3] *"How many blocks in a mile?"*. 2015. URL: http://www.quora.com/How-many-blocks-in-a-mile (visited on 07/17/2015).

[4] *"How many city blocks in a mile?"*. 2006. URL: https://answers.yahoo.com/question/index?qid=20060709191059AAwnmpf (visited on 07/19/2015).

[5] Metropolitan Police Department. *"DC Police Crime Mapping"*. N.d. URL: http://crimemap.dc.gov/ (visited on 06/22/2015).

[6] *"Data Download: Metrorail Ridership by Origin and Destination"*. 2012. URL: http://planitmetro.com/2012/10/31/data-download-metrorail-ridership-by-origin-and-destination/ (visited on 07/21/2015).

[7] *"Metro Station Entrances"*. 2015. URL: http://opendata.dc.gov/datasets/ab5661e1a4d74a338ee51cd9533ac787_50 (visited on 06/22/2015).

[8] *"Ward - 2012"*. 2015. URL: http://opendata.dc.gov/datasets/0ef47379cbae44e88267c01eaec2ff6e_31 (visited on 06/22/2015).